PHD

Group II Intron Thermophilic Reverse Transcriptases

Voina, Natasha

Award date:
2011

Awarding institution:
University of Bath

Group II Intron Thermophilic Reverse Transcriptases

Submitted by
Natasha June Voina

For the degree of Doctor of Philosophy
University of Bath
Department of Biology and Biochemistry
February 2011

ACKNOWLEDGEMENTS

I would like to thank my supervisors Professor Michael Danson and Dr David Hough for their guidance and the help that they have offered throughout my Ph.D. I am also incredibly grateful to Dr Duncan Clark as his endless enthusiasm and new ideas helped keep me motivated throughout the project and I am grateful for him allowing me the use of the laboratory facilities at GeneSys Ltd during the final year of my Ph.D.

I am very grateful to all the people in the University Lab 1.33 who helped make long days pass quickly and provided me with valuable knowledge as well as endless entertainment. These include Sarah Lamble, Charlie Nunn, Karl Payne, Nia Marrott, Philippe Mozzanega, Winnie Wu, Sylvain Royer, Jon Extance, Dalal Binjawar, Carolyn Williamson, Tracey Goult and Chris Vennard.

Additional thanks also go towards the staff at GeneSys Ltd who have always been very encouraging and survived all my experiments including the non lab-based baking variety. The staff at GeneSys are: Sarah Dugdale, Anna Siddle, Nick Morant, Deepal Pandya and Steve Millington

On a personal note I am eternally grateful for the love, support and encouragement that I have received from my family, especially my Mum and Dad, who have always encouraged me to do my best and be proud of what I achieve and to my Uncle Michael who has also offered continued support throughout my studies.

Special thanks also go to my sister, Suzanne Voina, who lived with me during the first year and has always been there for me during both the ups and downs. Her, if sometimes frustrating, rational thinking and ability to make me see past problems provided me with a sensible path to follow and taught me to relax and slow down.

A massive thank you will also go to Lee Hilton who has never wavered in his belief in what I can achieve. His interest in my project and his somewhat inaccurate description of what I do as ‘making tomatoes smile’ has reminded me not to take life too seriously and provided me with the strength I needed to see this through to the end and the determination that, one day, I will make them smile!

Group II Intron Thermophilic Reverse Transcriptases

CONTENTS PAGE

Group II Intron Thermophilic Reverse Transcriptases

ABSTRACT

A reverse transcription reaction allows the production of complementary DNA (cDNA) using an RNA template and relies on polymerases displaying reverse transcriptase (RT) activity. This process, with major applications in both research and in medical diagnostics, is often limited by the nature of the RTs available. RNA secondary structure can prove problematic where mesophilic retroviral RTs are used while the alternative approach, using thermophilic DNA polymerases with RT activity, often results in error-prone cDNA production.

This project recognised the need to study other possible sources of thermophilic RTs and outlines the study of four previously uncharacterised Group II Intron-encoded proteins (IEP), with RT domains, from thermophilic bacteria. While cloning of the IEP genes and their expression on a small scale proved successful, difficulties were encountered when attempting purification. Despite a lack of overall purity, samples containing IEPs from *Thermosinus carboxydivorans* and *Petrotoga mobilis* were shown to have RT activity but characterisation of these IEPs was not carried out. However, an IEP from *Bacillus caldovelox* proved to be an excellent candidate for characterisation as successful purification was achieved. Enzyme engineering was also performed, fusing a Sac7d domain onto the C-terminus of this protein. These enzymes were shown to have optimum RT activity at 54ºC with activity still being displayed at 76ºC. Other studies on these enzymes showed that, unlike the retroviral RTs, the IEPs displayed no DNA-dependent DNA polymerase activity. The Sac7d fusion protein was also studied in terms of possible enhancements to the RT activity of an IEP. However, preliminary studies showed that, although this domain did not prove to be detrimental to the enzyme, it had little effect on improving the processivity of the RTs.

Although this class of RT looks promising in terms of use as an alternative thermophilic RT, the IEPs studied in this report did incur major limitations during cDNA synthesis, which included lower than expected optimum reaction temperatures, very low fidelity and an inability to synthesise cDNA using complex RNA templates.

Group II Intron Thermophilic Reverse Transcriptases

ABBREVIATIONS

AMV	– Avian myeloblastosis virus
cDNA	– Complementary DNA
CHT	– Ceramic hydroxyapatite
CTAB	– Cetrimonium bromide
cv	– Column volume
DNA	– Deoxyribonucleic acid
dNTP	– Deoxyribonucleic acid triphosphate
DMSO	– Dimethyl sulfoxide
dsDNA	– Double-stranded DNA
DTT	– Dithiothreitol
EBS	– Exon binding site
gDNA	– Genomic DNA
HIV-1	– Human immunodeficiency virus type I
IBS	– Intron binding site
IEP	– Intron-encoded protein
IPTG	– Isopropyl β-D-1-thiogalactopyranoside
LB	– Luria Broth
LINE	– long interspersed nuclear elements
LTR	– Long terminal repeats
MMLV	– Moloney murine leukaemia virus
mRNA	– Messenger RNA
NEC	– No-enzyme control
ORF	– Open reading frame
PCR	– Polymerase chain reaction
PEI	– Polyethylenimine
qPCR	– Quantitative PCR
RE	– Restriction Endonucleases
RNA	– Ribonucleic acid
RNase H	– Ribonuclease H
rRNA	– Ribosomal RNA
RT	– Reverse transcriptase

RT-PCR – Reverse transcription polymerase chain reaction
SAP – Shrimp Alkaline Phosphatase
SDM – Site directed mutagenesis
SDS PAGE – Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SINE – Short interspersed nuclear elements
ssDNA – Single-stranded DNA
ssRNA – Single-stranded RNA
TSB – Tryptone soya broth
TB – Terrific broth
tRNA – Transfer RNA
X-Gal – 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside

Chapter 1 - Introduction

1.1 - INTRODUCTION

Discovery of Reverse Transcriptases

The Central Dogma theory was originally stated by Crick in 1958 and was covered again in a later article in 1970 and discussed the transfer of information between the three main polymers (Figure 1.1): DNA, RNA and protein. Three classes of transfers were considered.

- Class I – Transfers for which some form of evidence existed for their occurrence and included:
    - DNA-DNA – now termed replication and involves DNA-dependent DNA polymerases.
    - DNA-RNA – now termed transcription and requires a DNA-dependent RNA polymerase.
    - RNA-Protein – now termed translation and carried out by ribosomes.
    - RNA-RNA – This was presumed to occur due to the existence of RNA viruses and is now known to be carried out using RNA-dependent RNA polymerases.
- Class II – In 1958 no evidence and, at the time, no theoretical requirement could be seen for these transfers and therefore were considered rare or absent; this class contained:
    - DNA-Protein
    - RNA-DNA
- Class III – involved transfer from proteins. Due to the complexity of proteins, this class was removed from the central dogma diagram as they were considered impossible and included:
    - Protein-protein
    - Protein-DNA
    - Protein-RNA.

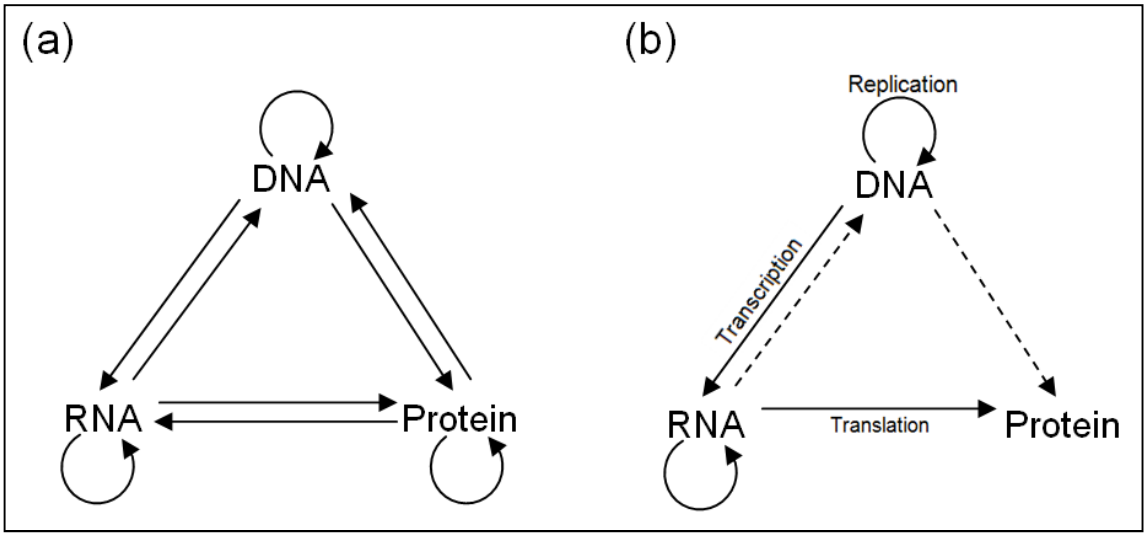

Figure 1.1:

(a) Represents all possible simple information transfers between the polymers.

(b) Represents the view of information transfers as stated in 1958.

In (b) the solid arrows represent all probably transfers while the dashed arrows represented possible transfers. The absence of arrows in (b) that are seen in (a) represents transfers thought to be impossible (Crick, 1970).

Work by two independent groups, Baltimore (1970) on Rauscher mouse Leukaemia Virus and Temin (1970) on Rous Sarcoma Virus, led to the discovery of the, once considered unlikely, transfer of information from RNA to DNA. The genomes of these viruses were known to be RNA and Temin and Baltimore independently showed that a DNA step was essential to their life-cycle by inhibiting infectivity with the use of DNA synthesis inhibitors. They concluded that, in order to carry out this process, a unique enzyme would be required displaying RNA-dependent DNA polymerase activity. This provided evidence for the Class II transfer of information of RNA-DNA and, due to the fact that the enzyme was functioning in reverse to regular transcription, the reaction was termed reverse transcription and the enzyme referred to as a Reverse Transcriptase (RT).

Sources of Reverse Transcriptases

Originally termed as unlikely or rare, the transfer of RNA-DNA is actually highly prevalent with a major involvement in shaping eukaryotic genomes, causing disease and also an essential requirement for the survival of many organisms. The main source of RT activity can be found in the form of retroelements. Retroelements are a class of mobile elements that, with the use of an RT and a DNA intermediate, move by a "copy and paste" mechanism allowing complementary DNA (cDNA) insertion into a host genome (Reviewed by Gogvadze and Buzdin, 2009). They replicate independently of the host genome using both cellular resources and element-encoded proteins, and are often considered 'selfish DNA' or nuclear parasites (Sabot and Schulman, 2006).

The only feature in common across all retroelements is their reverse transcriptase domain (Xiong and Eickbush, 1990), all of which contain seven amino acid motifs, totalling 178 residues, with important structural functions (Figure 1.2a). The deciphered structure of HIV-1 RT shows that, like other polymerases, RTs have a typical right-hand appearance including a palm, thumb and fingers. The location of the seven conserved motifs within all RTs can be seen within this diagram showing their location and structural importance (Figure 1.2b).

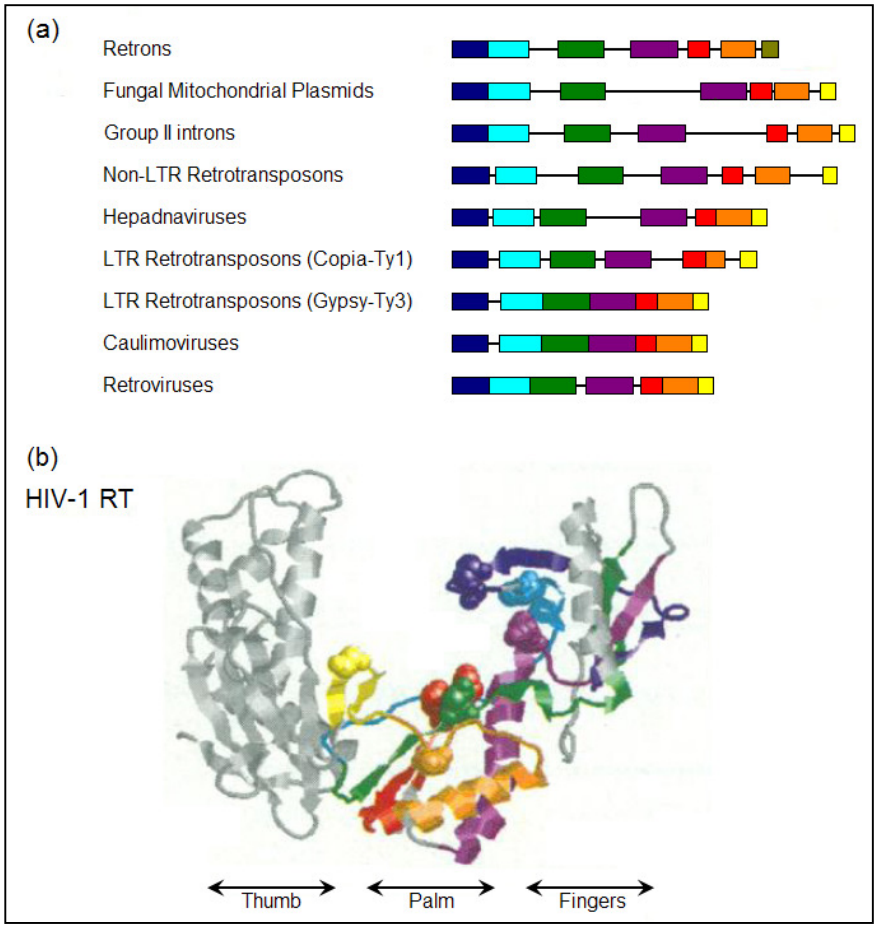

Figure 1.2:

(a) Represents the seven amino acid motifs conserved in all RTs from retroelements.

(b) The crystal structure of the p66 subunit of HIV-1 RT. A view from the back of the 'right-hand'.

The location of the motifs seen in colour in (a) can be seen on the HIV-1 RT structure (b) (modified from Nakamura *et al.* 1997).

Most structure and activity studies of RTs have been carried out on HIV-1 RT, but the basic mechanism of extending DNA by all polymerases is essentially the same. In the case of HIV-1 RT the polymerase binds to the template-primer substrate and the dNTP creating a ternary complex. A conformational change occurs, where the polymerase forms a 'closed' structure with the finger and thumb regions touching, which then orients the dNTP for catalysis. The palm of the polymerase is the active site containing aspartate residues and, with the use

of two metal ions, catalyses the incorporation of a new nucleotide on the 3'-OH of the primer. Another conformational change occurs as the polymerase forms an 'open' structure with the thumbs separating from the finger releasing the pyrophosphate that was liberated during the reaction (Castro *et al.* 2007, and reviewed by Herschorn and Hizi, 2010).

Origin of Reverse Transcriptases

RNA viruses, not including retroviruses, contain RNA as their genetic material. However, they do not use a DNA intermediate as part of their life cycle and therefore have no requirement for RT activity. With the fact that RNA viruses have a greater genomic diversity, both in organisation and sequences, and are present in a more diverse range of prokaryotes and eukaryotes than any retroelement, it is generally considered that they are in fact older than all the retroelements (Xiong and Eickbush, 1990). This, along with the fact that the seven amino acid motifs that define an RT can also be found in the RNA-dependent RNA polymerase of RNA viruses, can allow a phylogenetic tree to be constructed (Figure 1.3). This tree uses RNA virus RNA-dependent RNA polymerase as the root and draws relationships between the different types of retroelements. Rooting the tree in this manner is not making the assumption that RTs evolved from RNA-dependent RNA polymerase, but that they both shared a common ancestor.

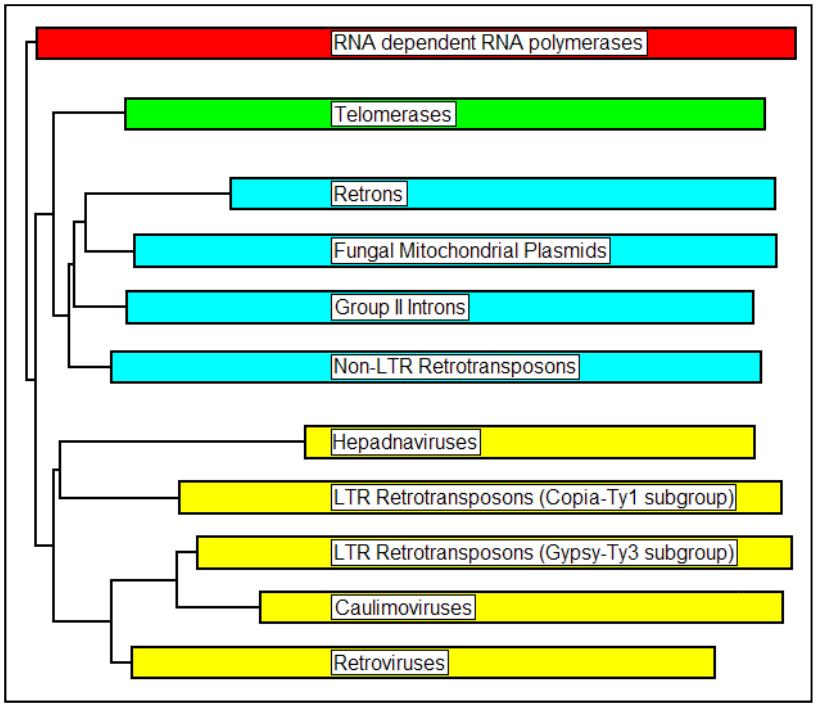

Figure 1.3: A possible phylogenetic tree of the relatedness of retroelements rooted with RNA dependent RNA polymerases from RNA viruses (red). The tree was constructed by Nakamura *et al*. (1997) using the neighbour joining method and the conserved domains of the RTs of the retroelements. Two main groups arise from this tree with the blue group representing the non-LTR retroelements and the yellow group representing the LTR retroelements. The RTs from telomerases are also included in the tree (green). The length of each box corresponds to the most divergent element within that box.

This phylogenetic tree puts the different retroelements into two main groups.

- The group highlighted in yellow are those that contain long terminal repeats (LTR) at both ends of their genomes and includes:
  - LTR retrotransposons
  - Hepadnaviruses
  - Caulimoviruses
  - Retroviruses
- The group highlighted in blue do not contain LTRs and includes:
  - Retrons
  - Fungal mitochondrial plasmids

- Non-LTR retrotransposons
- Group II Introns.

Role of Reverse Transcriptases in Nature

LTR Retroelements

LTR retroelements are characterised by the LTRs flanking their genomes formed during the tRNA primed RT step generating DNA from the RNA template (reviewed by Haig and Kazazian, 2004). Integration into the host genome does not occur in all LTR retroelements and where it does it is independent of the RT step.

- LTR retrotransposons – these elements constitute a large portion of eukaryotic genomes and have their ssRNA converted into DNA by their encoding RT. This DNA can then be integrated into a host genome by element-encoded integrase in a mechanism of transposition similar to that of retroviruses. Although very similar to retroviruses in that they contain a *gag* and a *pol* gene, the lack of an *env* gene, coding for an envelope glycoprotein means that their replication cycle is contained within a single cell and they are not infectious (as reviewed by Wilhelm and Wilhelm, 2001). There are two families of LTR retrotransposons which group according to the order of their domains in the *pol* gene:
  - Ty1-*copia* family are ordered protease-RT-integrase
  - Ty3-*gypsy* family are ordered protease-integrase-RT.
- Pararetroviruses – these are the Hepadnaviruses and Caulimoviruses that infect vertebrates and plants respectively. These elements have a partially single-stranded circular DNA genome that is 'filled in' by their RT upon infection to create closed-circular dsDNA. This completed genome can then be transcribed into RNA and translated into virus encoded proteins. Unlike retroviruses, these viruses do not integrate into the host genome (reviewed by Flavell, 1995)

- Retroviruses – these are viruses that require RT activity to allow their ssRNA genome to be converted into dsDNA allowing its insertion into a host genome. Unlike LTR retrotransposons, the presence of an *env* gene within their genome allows the expression of an envelope glycoprotein enabling the retrovirus to enter a new host cell. The RT of retroviruses will be discussed in greater detail further on in this chapter.

Non-LTR retroelements

These elements lack flanking LTRs and instead typically contain a 3' poly(A) tail and the 5' end often has various truncations. The RT step occurs at the genomic target site in a process termed target-primed reverse transcription, allowing integration of the element into the host genome (reviewed by Haig and Kazazian, 2004).

- Retrons – as yet have no known function but are responsible for the production of multi-copy small ssDNA (msDNA) in bacteria. As reviewed by Travisano and Inouye (1995), the retron consists of an RT gene (*ret*) and at least the two other genes *msr* and *msd*. Once an RNA transcript is produced, the RT is translated and is responsible for the production of cDNA from msd RNA. The final msDNA element contains a chimera of ssDNA linked to ssRNA.
- Fungal mitochondrial plasmids – found in species of *Neurospora* by Collins *et al*. (1981) these retroelements, of which Mauriceville and Varkud are the best characterised, are small double-stranded plasmid DNA found in mitochondria. As reviewed by Griffiths (1995), these plasmids have been shown to insert into DNA via an RNA intermediate. They replicate via a transcription-reverse transcription method and utilise a mitochondrial RNase to remove the RNA template.
- Non-LTR retrotransposons – these are one of the most abundant transposable elements of which there are two types:

    - Long interspersed nuclear elements (LINEs) – The LINE's 'copy and paste' replication mechanism means they occupy approximately 20% of most mammalian genomes. The full-length mRNA transcript contains two ORFs with ORF1 coding for nucleic acid chaperones that form trimers creating an RNP with the LINE mRNA. ORF2 encodes an endonuclease and an RT required for retrotransposition (reviewed by Belancio *et al*. 2008).
    - Short interspersed nuclear elements (SINEs) – SINES are considered non-autonomous retroelements since they rely on ORF2 of LINEs for their retrotransposition.
- Group II Introns – found in the organelles of lower eukaryotes, in bacteria and in the *Methanosarcina* genus of Archaea (Toro, 2003). These mobile elements are able to interrupt genes with no deleterious effect to the organism due to their ability to auto-splice from the mRNA transcript. Once spliced out of the transcript, the reverse transcriptase is responsible for generating a DNA copy of the intron allowing its insertion into the genome. It has been theorised that Group II Introns are in fact progenitors of nuclear spliceosomal introns due to the similar nature of their splicing. Group II Introns will be discussed in more detail towards to the end of this chapter.

Telomerases

Since DNA polymerases cannot complete the synthesis of both strands of blunt ended DNA during replication, the role of a telomerase is to protect the ends of linear chromosomes. Forming a ribonucleoprotein (RNP), the RNA acts as a template for the synthesis of the telomeres with the use of the telomerase reverse transcriptase (TERT) (as reviewed by Nakamura and Cech, 1998). The seven previously mentioned RT motifs have been found in TERT, showing that it is both functionally and structurally related to retroelement RTs. The inclusion of TERT within a phylogenetic tree proved difficult as, unlike other RTs,

telomerases are not transposable elements and are therefore not replicated by the low fidelity RTs. However, by rooting the tree with RNA-dependent RNA polymerases as the common ancestor, the telomerases were found to be most closely related to the group containing the non-LTR retrotransposons (Eickbush, 1997; Nakamura *et al*. 1997).

Applications of RTs

The discovery of RT activity has proved essential in the cloning and expression of eukaryotic genes using bacterial systems. As reviewed by Sharp (1994) eukaryotic genes are often interrupted by large regions of non-coding sequences called introns. During eukaryotic protein expression, these introns are removed from the RNA transcript in the nucleus, in a process called splicing, before the messenger RNA (mRNA) is directed to the cytoplasm where translation occurs. Bacteria lack this machinery to splice introns from an RNA transcript. Therefore, cloning a gene into bacteria directly from eukaryotic genomic DNA would not produce the correct protein. Creating cDNA directly from an mRNA transcript offers a huge advantage as it eliminates the need of intron splicing. Utilising RT activity has therefore made it possible to express eukaryotic proteins in bacterial systems.

Identification of individual mRNA transcripts also indicates the proteins that are being expressed at that particular time within a cell or tissue sample (Richert *et al*. 1996, Gravelat *et al*. 2008). Studying mRNA allows gene expression to be monitored and, coupled with qPCR, allows the mRNA levels within the sample to be quantified. Gene expression studies provide a useful tool for a variety of different applications including drug trial analysis, identification of developmental pathways and assessing the affect of environmental conditions on a particular tissue type or organism. Care has to be taken with this technique as it has been documented that mRNA levels within a cell do not necessarily correlate to

protein levels within the cell (Gygi *et al*. 1999). However, this technique still forms a necessary tool for the study of gene expression.

RTs have also proved to be important in disease diagnosis. RT-PCR and RT-qPCR are essential as a quick and sensitive method for diagnosis of infectious human diseases (Towner *et al*. 2004), live stock diseases (Moniwa *et al*. 2007) and also in monitoring certain types of cancer (Takano *et al.* 2000). The speed and sensitivity of this method are often required in highly contagious diseases, where early diagnosis can help to quarantine individual cases and contain the disease.

Commercially Available Reverse Transcriptases

The main source of RTs available commercially is retroviruses. As reviewed by Herschorn and Hizi (2010), RTs are essential to the life cycle of a retrovirus. These viruses have a genome consisting of two copies of ssRNA and require an enzyme to convert their RNA genome into dsDNA that can be inserted into the host genome. The RT enzyme they possess therefore has three activities:

- RT activity to create a cDNA strand
- RNase H activity to degrade the RNA in the DNA:RNA hybrid
- DNA-dependent DNA polymerase activity to create the dsDNA.

The RNA- and DNA-dependent DNA polymerase activities are carried out in the same active site. However, the RNase H activity requires an additional active site and is joined by a connecting region to the typical right-hand polymerase conformation. The main two retroviral RTs that are used for cDNA synthesis are the monomeric retroviral RT isolated from Moloney Murine Leukaemia Virus (MMLV-RT) (Verma, 1975) and the heterodimeric RT isolated from Avian Myeloblastosis Virus (AMV-RT) (Houts *et* al. 1979). The heterodimer of AMV-RT is made up of two polypeptide subunits, α and β, with molecular weights of 65kD and 105kD respectively. The smaller α subunit is structurally related to

the β subunit and results from proteolytic cleavage of the β subunit (Rho *et al*. 1975). A typical RT reaction with the use of these mesophilic RTs can be seen in Figure 1.4.

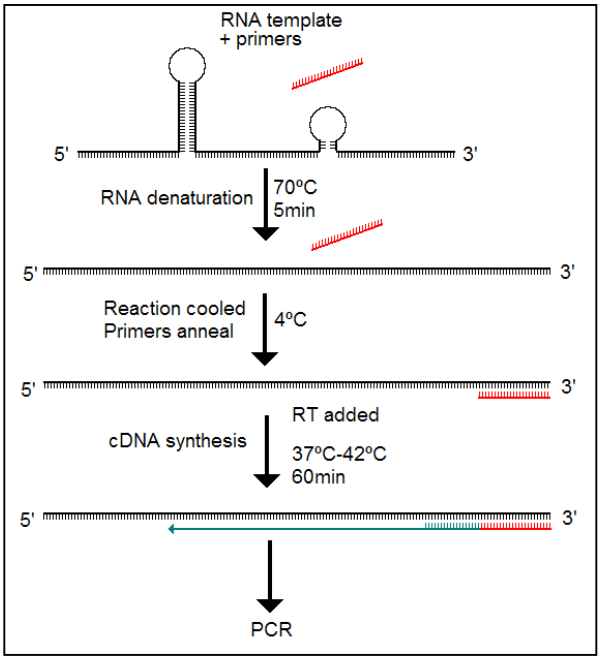

Figure 1.4: A typical RT reaction. RNA is initially heated, allowing denaturation of any secondary structure. The reaction is then cooled, allowing primers to anneal and the RT to be added. cDNA synthesis occurs when the reaction is incubated, typically for 60min at 37-42℃. The cDNA produced can then be used for further applications.

As an alternative to the retroviral RTs, thermophilic DNA-dependent DNA polymerases, typically used for PCR, can also show reverse transcriptase activity with the use of manganese as the divalent cation instead of magnesium (Myers and Gelfand, 1991). The thermophilic nature of the enzyme prevents the need to cool the reaction before the addition of the RT and prevents RNA secondary structure becoming detrimental to the cDNA yield as will be discussed below. Subsequent cDNA amplification can also be carried out with the same enzyme, therefore eliminating the need for additional enzymes or reactions steps.

Limitations of the Current RTs

Although a well established technique, cDNA synthesis does incur several disadvantages, both as a fault of the enzymes used and the difficulty of the RNA template that is essential for the reaction.

Low Fidelity

Many DNA-dependent DNA polymerases possess a 5'-3' exonuclease activity. However, this is lacking in retroviral RTs, possibly as a method to allow the virus to continue evolving and evading immune responses. HIV-1 RT has been reported to mis-incorporate 1 nucleotide in every 6,900 nucleotides polymerised, while MMLV-RT has higher fidelity with a mis-incorporation rate of 1 in 28,000 (Ji and Loeb. 1992). The lack of this proof-reading ability can lead to mutations being incorporated into the cDNA, which becomes problematic when the cDNA is required for downstream applications such as cloning and sequencing, where an accurate representation of the RNA is required (Malboeuf *et al*. 2001).

Low fidelity can also be found when using thermophilic DNA polymerases possessing RT activity. The presence of manganese has been documented to lower the fidelity of polymerases (Goodman *et al*. 1983). The manganese can lower the specificity of the 3'-5' exonuclease, which normally removes unpaired and mismatched nucleotides at the primer terminus prior to elongation, and also reduces the accuracy of base-pair selection (El-Diery *et al*. 1984). Therefore, cDNA synthesis with these enzymes is often error prone and any manganese that contaminates subsequent PCRs may also prove to be mutagenic.

Low Yields

A common problem with the wild-type mesophilic RTs is the low yield of full-length cDNAs produced and a background level of truncated products. Different

types of mRNA templates can be reverse transcribed with different efficiencies (Buell 1978) which can be attributed to two main factors: the mesophilic nature of the enzymes and the RNase H activity associated with the RT.

Mesophilic RTs

The typical RT reaction temperature used *in vitro* for cDNA synthesis is between 37-42ºC for wild-type mesophilic retroviral RTs. RNA is single-stranded and therefore often forms complex secondary structures, such as loops and hairpins. Although this secondary structure can be denatured prior to enzyme addition, the low reaction temperatures can allow this to reform. Studie*s* have shown that HIV-1 RT makes contact with the RNA spanning a region of approximately 30 bases at any one time. Small hairpins can stall the RT when they reach the active site of the enzyme, while a larger hairpin can pause the RT when the finger domain of the enzyme, spanning approximately 7 nucleotides, makes contact with the hairpin before it can reach the catalytic core of the RT (Figure 1.5) (Harrison *et al.* 1998). The RT is then in contact with the template for a further 23 bases after its catalytic core, during which secondary structure within this region can still act to pause the polymerase (Harrison *et al. 1*998). When an RT does pause, it can dissociate from the template and so cease the synthesis of full-length cDNA. Therefore, sites of secondary structure within RNA templates, both 5' and 3' of the catalytic core, can interfere with efficient cDNA synthesis, allowing the accumulation of truncated products and low yields of the full-length cDNA.

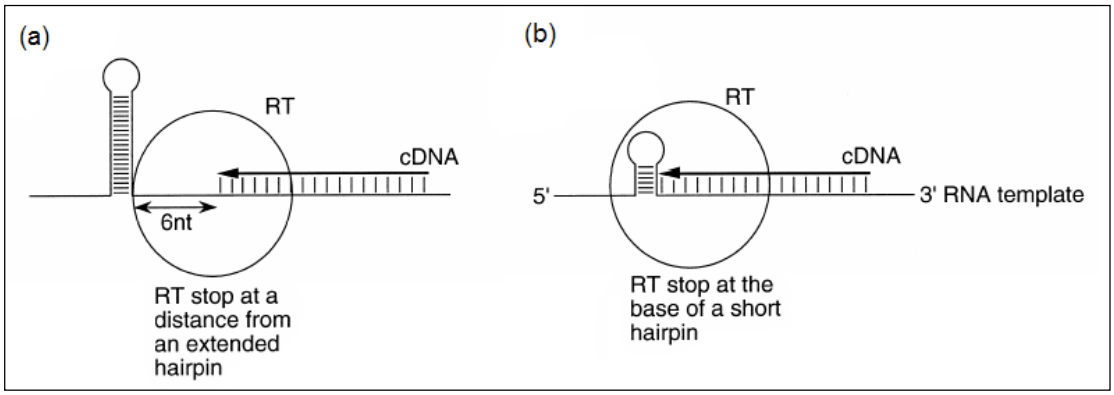

Figure 1.5: Effects of secondary structure of RNA on an RT. HIV-1 RT spans a region of approximately 30 nucleotides. RNA secondary structure present within this region can pause the RT, with large hairpins pausing the RT 6 nucleotides before it reaches the catalytic core as it interacts with the finger domain of the enzyme (a) and smaller hairpins stalling the RT as they reach the catalytic core (b). (figure from Klasens *et al.* 1999).

RNase H activity

Retroviral RTs require RNase H activity to degrade the RNA strand of an RNA:DNA hybrid prior to the creation of dsDNA. However, *in vitro* this activity can limit the yield of full-length cDNA synthesis. The RNase H activity is an endonuclease that cleaves the RNA at the 3'-OH end of the growing DNA chain. This allows un-copied RNA to dissociate, preventing further cDNA extension, and removes a portion of the mRNA for further cDNA synthesis (Kotewitcz 1988). This problem is further increased by the pausing of the RT at regions of secondary structure, as mentioned above, as this increases the likelihood of a cleavage event (Gerard 1997).

Low Primer Specificity

A high background signal of non-specific products can be produced when using mesophilic RTs with specific primers. The 42ºC reaction temperature is normally below that of the melting temperature of the primers, and therefore false annealing can occur (Freeman *et al*. 1996).

Current Solutions to Reverse Transcriptase Limitations

One of the main solutions to the problems mentioned above has been to manipulate the wild-type RTs in the hope of enhancing the enzymes. Two of the main manipulations have been to increase the thermostability of the enzyme and to remove the RNase H activity.

Thermostable Reverse Transcriptases

A thermostable enzyme will allow cDNA synthesis to occur at greater efficiency. The higher temperature will allow secondary structures within the RNA template to be denatured and will also increase the specificity of the synthesis when using gene-specific primers (Fuchs *et al*. 1999). Freeman *et al*. (1996) discuss unpublished work by Xu and Sonntag where they noticed a decrease in background signal of RT-PCR when increasing the reaction from 42 to 45ºC, showing that just a small increase of 3ºC can greatly improve the efficiency of the reaction. Utilising thermophilic DNA-dependent DNA polymerases where mRNA secondary structure becomes problematic is one solution to increasing cDNA yields. However, the presence of manganese causes additional problems as mentioned previously. Manufacturers have mutated mesophilic retroviral RTs to increase their thermostability compared to the wild-type enzyme. ThermoScript™, from Invitrogen, is an AMV-RT with point mutations in the RNase H domain that reduce RNase H activity and also enhance its thermostability. The optimum reaction temperature of ThermoScript is 65ºC, compared to the wild-type AMV-RT that has a recommended reaction temperature of 37ºC. ThermoScript™ will synthesise cDNA at temperatures as high as 70ºC (Schwabe *et al*. 1998). SuperScript™III, a mutated MMLV-RT also manufactured by Invitrogen, is reported to be fully active at 50ºC. Both enzymes are marketed as having high performance with difficult RNA templates with the ability to achieve longer cDNA lengths and higher yields.

RNase H Minus Mutants

Due to the fact that the DNA polymerase and RNase H activities of retroviral RTs are structurally separated and their activities are not coupled (DeStafano *et al*. 1991), RNase H minus mutants can be generated. Kotewicz *et al*. (1988) found that the amino acids 503-611 of MMLV-RT are required for RNase H activity. However, deletion of one fourth of the MMLV-RT, to remove this activity, prevented the enzyme from synthesising cDNA efficiently. Manufacturers now offer modified retroviral RTs that contain point mutations to reduce RNase H activity. These mutations are found both in AMV-RT, Invitrogen's ThermoScript™ (Schwabe *et al*. 1998), and MMLV-RT, SuperScript™ also from Invitrogen. These mutant enzymes are marketed by their manufacturers as able to synthesise more full-length cDNA due to the reduction in RNase H activity helping to eliminated RNA degradation during the course of the reaction.

As an alternative to manipulating the RT enzymes themselves, other attempts to improve RT reactions often involve additions or alterations to the typical reverse transcription protocol.

Nucleocapsid proteins

Nucleocapsid (NC) proteins have been used to try to reduce the limitations of RNA secondary structure without the need for increasing temperatures of the RT reaction. NC proteins are small basic proteins that have a structural role in the retroviral lifecycle, promoting efficient DNA synthesis during replication (as reviewed by Darlix, 1995). The NC protein is thought to aid in reverse transcription by destabilising regions of secondary structure within the RNA. This ability was utilised by Wu *et al.* (1996), who showed that the presence of HIV-1 NC protein could reduce pausing of the MMLV-RT 8-10 fold during cDNA synthesis. However, Das *et al*. (2001) found that improvements in the quality of

first-strand synthesis reactions with the addition of HIV-1 NC protein were only seen on the RNase H minus variants with no improvement on wild-type MMLV-RT. Since RNase H minus mutants tend to have improved thermostability this could negate the requirement for the addition of NC to the reaction.

An Alternative Solution – A Naturally Occurring Thermophilic RT

Thermophiles belong to a larger group of organisms known as extremophiles. This group includes organisms from all domains of life that live in environments that are regarded as extreme. These conditions can include extremes such as temperature, radiation, pressure, desiccation, salinity and pH (Rothschild and Mancinelli 2001) (Table 1.1).

Proteins from thermophilic organisms need to have structures adapted to allow their correct folding and efficient functioning at the high growth temperatures of their native organism. These proteins could overcome these extremes in temperature by either lowering their rate of unfolding at the high temperatures and/or increase their rate of (re)folding at these conditions. All studies so far link the higher thermodynamic stability of thermophilic proteins to a decrease in their rate of unfolding at the higher temperatures (Das and Gerstein. 2001).

Das and Gerstein (2001) reviewed the different methods that thermophilic proteins adopted in order to maintain their thermostability and ordered them into two main categories:

1. Stabilising features considered to be of minor importance, fond in only a certain number of families and included:
    I) An increased number of hydrogen bonds
    II) Higher packing density with smaller more condensed proteins

III) Shorter polypeptide chains
IV) Improved hydrophobic interactions
V) Optimised surface area

2. Major stabilising features that have been found in most thermophilic protein studies
    I) Increase in electrostatic interactions by formation of salt bridge networks.
    II) Stabilisation of α-helices
    III) Increase in the fraction of Pro and β-branched amino acids decreasing the entropy of the unfolded state
    IV) Decrease in uncharged polar residues which helps avoid chemical degradation at the high temperatures

A combination of several of these methods will render a thermophilic protein more thermostable than its mesophilic homologue despite the fact that, on initial analysis, the two proteins would seem to have a very similar structure.

Environmental Parameter	Extremophile	Definition	Examples
Temperature	Hyperthermophile	Growth >80ºC	*Pyrococcus furiosus*
	Thermophile	Growth 60-80ºC	*Thermus aquaticus*
	Pyschrophile	Growth <15ºC	*Psychrobacter*, some insects
Radiation			*Deinococcus radiodurans*
Desiccation	Xerophiles	Anhydrobiotic	*Arternia salina*; nematodes, fungi, lichens
Salinity	Halophile	2-5M NaCl	Halobacteriaceae, *Dunaliella salina*
pH	Alkaliphile	pH>9	*Natronobacterium, Bacillus firmus* OF4,
	Acidophile	pH<4	*Cyanidium caldarium, Ferroplasma* species
Oxygen Tension	Anaerobe	Cannot tolerate oxygen	*Methanococcus jannaschii*

Table 1.1: Classification and examples of extremophiles (adapted from Rothschild and Mancinelli, 2001).

Industrial processes can be expensive and non-specific, which can result in the production of unwanted biproducts that will need to be removed and safely disposed of. The use of mesophilic enzymes has been limited, due to their restrictions at extreme environments limiting their use in an industrial setting (Demirjian *et al*. 2001). The discovery of extremophilic organisms has allowed their exploitation in situations where their mesophilic counterparts would not survive. These extremophiles often display novel metabolic pathways, and the

organisms and their enzymes can be utilised in situations that can combine biochemistry with industrial processes (Hough and Danson, 1999).

An example of thermophilic enzymes revolutionising an already established scientific research technique can be seen in their use in DNA amplification. When PCR was first invented, its major drawback was the use of the Klenow fragment of *E. coli* DNA polymerase I (Saiki *et al*. 1985). The template denaturation step during the PCR also caused the thermolabile enzyme to be destroyed, and therefore fresh enzyme had to be added at every extension step in the cycle. As a consequence, PCR was labour intensive, expensive and not achievable on a large-scale. The replacement of the Klenow fragment with the *Thermus aquaticus* (*Taq*) DNA polymerase revolutionised this technique. This polymerase is heat resistant, and therefore not denatured in the high temperature template denaturation step. Using thermophilic enzymes simplified this technique and also improved the actual performance, with enhancements seen in yield, sensitivity and the length of targets that could be amplified (Saiki *et al*. 1988)

Group II Introns

A novel potential source of thermophilic RTs was recognised by Vellore *et al.* (2004) and Ng *et al.* (2007). Their work involved studying intron-encoded proteins (IEPs), which posses an RT domain, from a Group II Intron in thermophilic bacteria. They showed that these proteins did indeed possess RT activity at high reaction temperatures compared to that of the wild-type retroviral RTs. The work by these two groups will be discussed in greater detail throughout this report.

As mentioned previously, Group II Introns are found in all three domains of life both in non-coding and coding regions of the genome. These elements are found in mitochondria and chloroplasts of eukaryotes but have not yet been discovered in the nuclear genomes of these organisms. In bacteria the introns tend to be found in non-coding regions or on other mobile elements such as plasmids in copies of one to a few per cell (Bechocine *et al*. 2007). They do not seem to confer any advantage to their host organism however; providing they retain the ability to self-splice they can interrupt genes with no deleterious effects.

Intron Structure

The actual DNA sequences of Group II Introns tend to be very diverse. However, they are defined by a conserved RNA secondary structure that consists of six double-helical domains radiating from a central wheel (Figure 1.6). This structure forms a catalytically-active ribozyme capable of splicing itself from an RNA transcript. Fedorova and Xingler (2007) reviewed the different domains in the intron structure, which include:

- Domain I – the largest of the functional domains and essential for splicing. This domain contains two 5' exonic substrate recognition sequences (EBS1 and EBS2) that interact with the corresponding sequence on the 3' end of the 5' exon (IBS1 and IBS2). These interactions help identify the 5' splice site and also aid in the recognition of specific sites for homing to occur. Domain I also contains the EBS3, which is essential for recognition of the 3' splice site (IBS3) required in the second step of the splicing reaction. Domain I is also responsible for forming a scaffold allowing the assembly of the other domains into the catalytic structure.
- Domain II – contains sites for interactions with Domain I and Domain VI forming long-range tertiary contacts aiding in the structure of the ribozyme.

- Domain III – referred to as a catalytic effector. It is expected that this domain forms tertiary interactions with the intron and its presence enhances reaction rates of Group II Introns. However, it is not strictly required for catalysis
- Domain IV – resides outside of the catalytic core of the ribozyme and is highly variable in length. It can contain an ORF that codes for an IEP that will be discussed in further detail below.
- Domain V – along with domain I this domain is also considered absolutely essential for splicing and is the most phylogenetically conserved of the whole intron. It has a structural role with its binding to domain I as well as a role in catalysis
- Domain VI – contains the bulged (non base paired) adenosine that is essential for the splicing reaction, which is discussed in detail below.

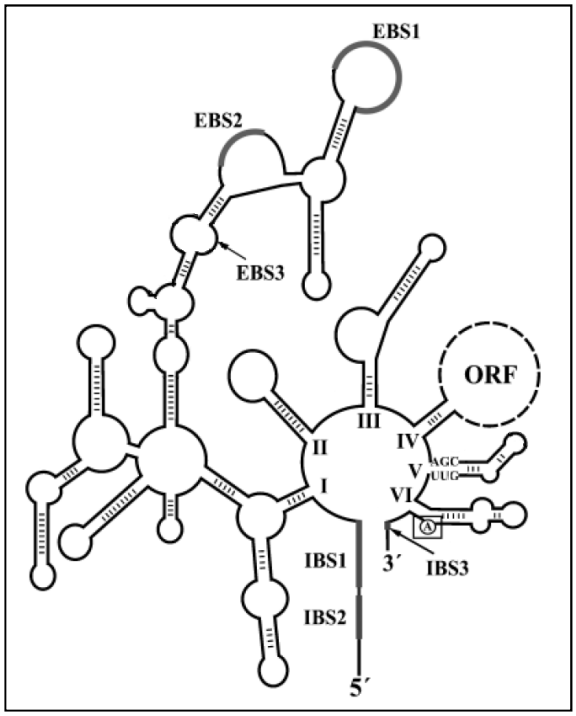

Figure 1.6: Generic RNA secondary structure of a Group II Intron forming a ribozyme. Domain IV contains an ORF with possible RT activity and resides outside the catalytic core of the ribozyme (Toro *et al.* 2007).

Intron-Encoded Protein

An ORF is often found in domain IV which resides outside the catalytic core of the ribozyme (Figure 1.6). This single ORF codes for an intron-encoded protein that possess up to four different enzymatic activities. These activities aid in the mobility and the splicing of the intron. In bacterial Group II Introns this IEP consistently contains an RT domain, although in some cases this is presumed inactive due to the loss of some of the conserved motifs (Mohr *et al*. 1993). This RT domain is then followed by domain X. Domain X has been designated as a maturase as it is thought to stabilise the catalytically-active RNA structure (Blocker *et al*. 2004). The predicted secondary structure of the LtrA IEP from *L. lactis* has been compared with the crystal structure of HIV-1 RT (Figure 1.7). This comparison shows that the RT domain of the IEP contains the same seven conserved motifs that are found in HIV-1 RT and all other RTs. The comparison also showed that the domain X appears to have retroviral RT structural features that relate to the 'thumb' and connection domains of the polymerase (Blocker *et al*. 2007). Since these domains in retroviral RTs bind the template RNA and primer, it has been speculated that the maturase domain functions in specific binding of the intron RNA for splicing and reverse transcription (Mohr *et al*. 1993). Even in situations where the RT motifs of the IEP are mutated, and presumably have lost their function, the maturase domain tends to remain intact suggesting it has an essential role in the splicing of the intron (Mohr *et al*. 1993).

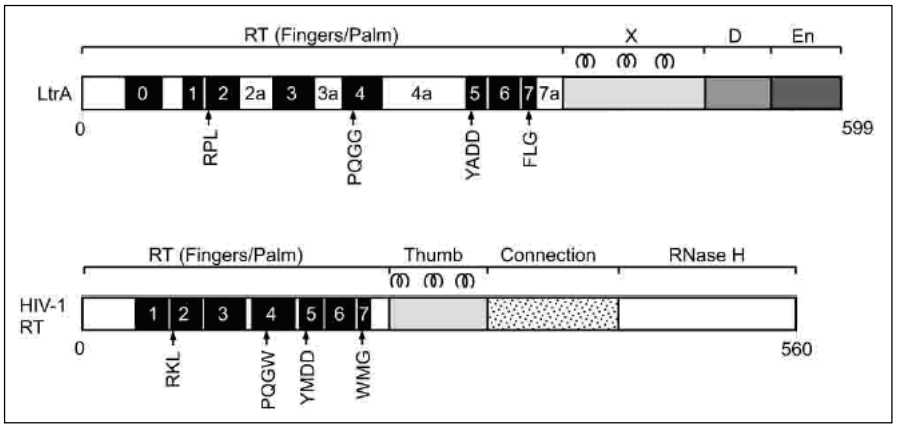

Figure 1.7: LtrA (IEP) from *Lactococcus* compared with HIV-1 RT. Domain 0 is characteristic of RTs from non-LTR-retroelements. Domains 1-7 are conserved sequences found in all RTs in the finger and palm domains; the actual conserved motifs are shown below. RNA-binding domain (X), containing 3 alpha helices, correspond to those found in the HIV-1 RT thumb domain. (Figure from Blocker *et al.* 2007).

While some IEPs from bacteria only contain the RT and maturase domains, others have been associated with C-terminal DNA-binding and/or DNA endonuclease domains (Figure 1.8). The endonuclease domain can aid re-insertion of the intron by cleavage of the antisense strand of the target DNA (Filippo and Lambowitz. 2002). One theory suggests that the earliest bacterial introns contained just the RT and the maturase domains, and therefore evolved mobility in the absence of this endonuclease. It is thought that the endonuclease was therefore acquired separately from a bacterial family of DNA nucleases (Zimmerley *et al.* 2001). However, another theory suggests that these domains have in fact been lost from the IEPs that now only contain the RT and the maturase domains (Mohr *et al*. 1993).

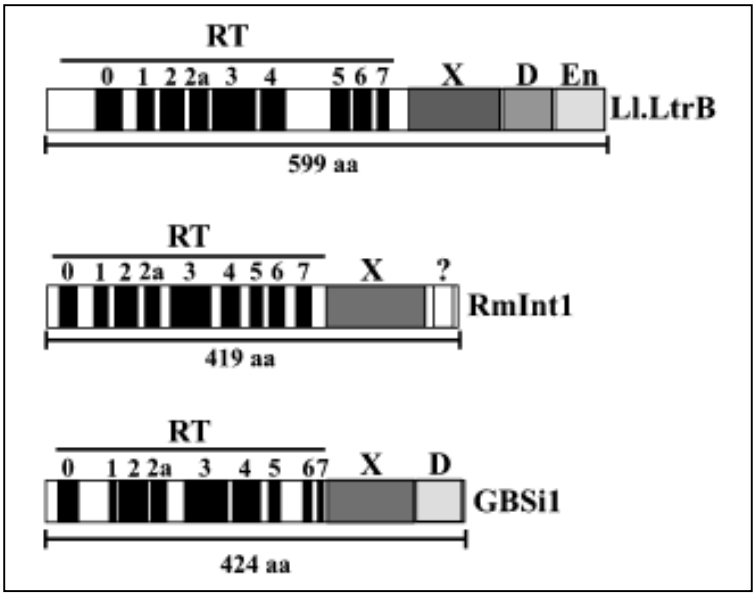

Figure 1.8: A model showing the various domains that can be found within bacterial IEPs. Shaded, numbered areas in the RT domain correspond to conserved domains found across RTs (Blocker *et al*. 2004). IEP Ll.LtrB found in *Lactococcus lactis* contains all 4 domains, RmInt1 from *Sinorhizobium meliloti* is lacking both the DNA-binding (D) and the DNA endonuclease (En) domains and GBSi1 from group B *Streptococcus* species is lacking the endonuclease domain but the DNA-binding domain is present. Diagram taken from Toro *et al*. (2007).

Splicing Mechanism

In order to prevent permanent interruption of an RNA transcript, and therefore incorrect protein expression, it is necessary for the Group II Intron to splice from the RNA. This relies on the conserved secondary structure of the RNA and the maturase that is coded for by domain IV (Wank *et al*. 1999). Where the IEP is lacking from the intron it is assumed that host proteins must be relied on for splicing to occur. The recognition of the 5' splice site occurs with the interaction of EBS1 and EBS2 of the intron with the IBS1 and IBS2 of the exon, and the 3' splice site is recognised by interaction of EBS3 with IBS3 (Figure 1.6). The bulged adenosine found in domain VI contains an exposed 2'-hydroxyl group that acts as nucleophile attacking the phosphate at the 5'-end of the intron. An intermediate product forms where Exon 1 is released and the intron lariat is still

connected to Exon 2. Base pairing interactions of Exon 1 to the intronic RNA still occur, which position the 3'-OH site of Exon 1 to allow a second nucleophilic attack at the 3' splice site. During this step the two Exons are joined together and the lariat RNA is released (Saldanha *et al.* 1993; Fedorova and Zingler *et al*. 2007) (Figure 1.9).

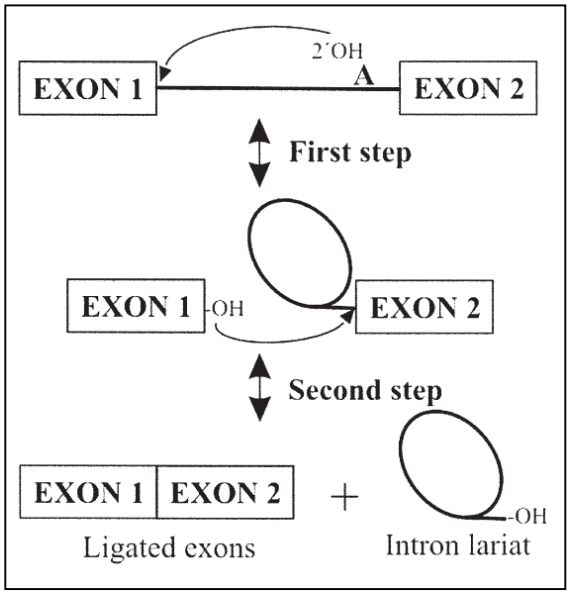

Figure 1.9: Splicing of the intron from an RNA transcript via a lariat intermediate. The 2'-OH of the bulged adenosine attacks the phosphate at the 5' splice site releasing Exon 1. The 3'-OH group of Exon 1 acts as a second nucleophile attacking the 3' splice site. The final product consists of the two exons ligated together and a released intron lariat. Diagram from Toro (2003).

Mobility

Group II Introns are mobile elements that have two types of mobility:

Retrohoming

Retrohoming occurs when the Group II Intron copies itself into an identical but intron-less DNA site, a trait that requires some form of recognition sequence. It is thought that in LI.LtrB integration EBS1 of the intron base pairs to the IBS1

site 5' to the intron insertion site (Cousineau *et al.* 2000). The RNP consisting of the intron RNA and the IEP promotes reverse splicing of the intron into the dsDNA and endonucleolytic cleavage. As the RNA splices itself into the sense strand the second strand is cleaved by the endonuclease which allows this strand to act as the primer for the cDNA synthesis of the intron (Cousineau *et al*. 2000; Martínez-Abarca *et al*. 2000; Moher *et al.* 1993; Fedorova and Zingler *et al.* 2007). This cleavage event is unlikely to be essential as homing has been reported in introns lacking the endonuclease domain (Martínez-Abarca *et al*. 2000), although the mechanism for retrohoming in the absence of this activity is unclear it is known to be Rec-A independent, unlike retrotransposition, and likely to occur at a replication fork or a site of an actively transcribe gene. The RT activity generates cDNA, which is then followed by second-strand synthesis followed by repair (Figure 1.10a).

Retrotransposition

Retrotransposition is when the intron inserts into an entirely new site and occurs less frequently than retrohoming. This event is thought to be independent of the endonuclease activity and instead involves host-encoded RecA. Instead of reverse splicing into a specific dsDNA site, the intron reverse splices into an exon RNA. cDNA synthesis occurs generating a DNA strand containing the exon and the Group II Intron, and this is followed by second-strand synthesis. In order to integrate into the genome, a homologous recombination event has to occur, which is dependent on the host encoded RecA, and then resolution leaves a once intron-less chromosomal allele becoming an intron-containing allele (Cousineau *et al*. 2000) (Figure 1.10b).

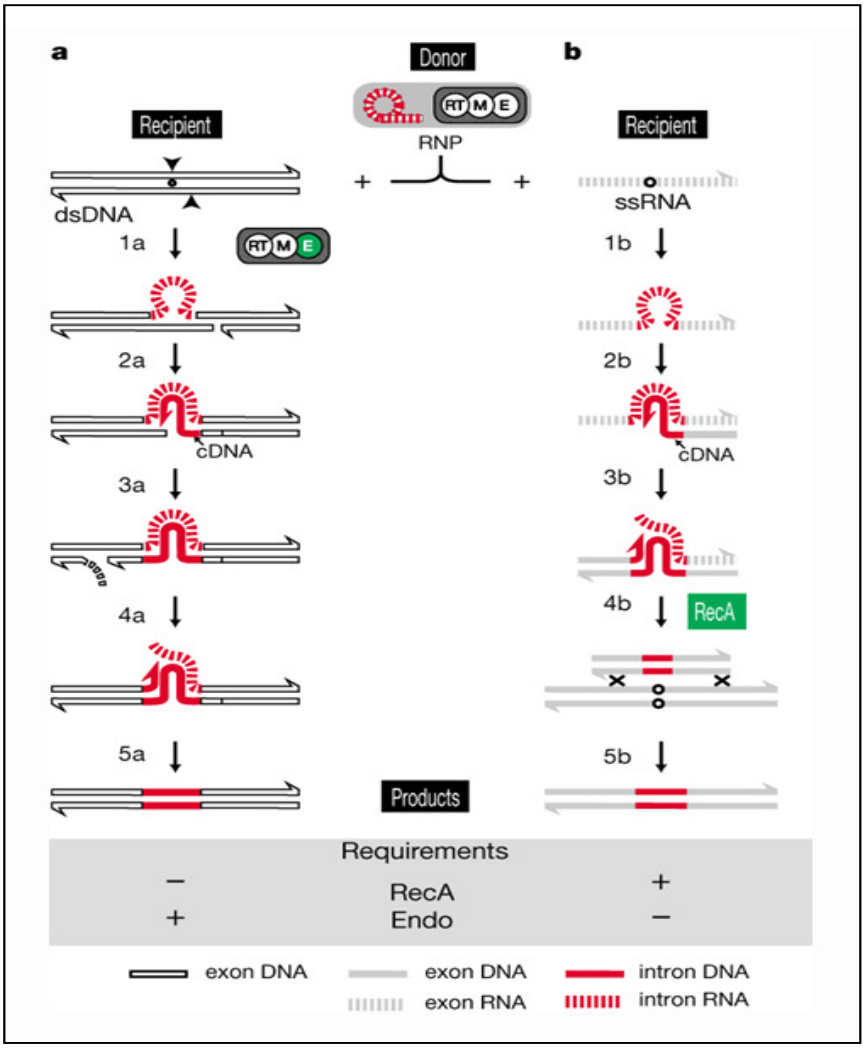

Figure 1.10: A diagram showing a) the retrohoming and b) the retrotransposition of the LI.LtrB Group II Intron. The RNP is depicted on a grey background with the intron lariat (red) and the trifunctional protein, LtrA with reverse transcriptase (RT), endonuclease (E) and maturase (M) activities. Functions required exclusively for one pathway are shown in green.

a) Retrohoming occurs when the intron reverse splices into dsDNA and endonuclease cleavage occurs (1a) then cDNA synthesis of the intron and the DNA occurs (2a and 3a) followed by second-strand synthesis (4a) and finally repair (5a).
b) Retrotransposition occurs when the intron reverse splices into RNA (1b) followed by cDNA synthesis (2b) and then second-strand synthesis (3b). A recombination event then has to occur to integrate the dsDNA containing the intron into the genome (4b), followed by resolution creating an intron-containing allele (5b).

(Cousineau *et al*. 2000)

Aims of Project

The aim of this Ph.D project was initially to isolate bacteria, from environmental soil samples, that grew at temperatures above 60ºC. A screening process would then be used to identify any bacteria that contained an IEP RT domain from a Group II Intron. In parallel to this screening process, a bioinformatics approach would also be adopted to identify any sequenced but uncharacterised IEP from thermophilic bacteria.

Identification and cloning of these IEP genes would follow the screening process, along with the expression and purification of the protein. Additional manipulations could be made during the cloning stage of the IEP gene to generate the fusion of protein domains that could potentially enhance the properties of the enzyme.

Once the proteins had been successfully expressed and purified, research would begin on defining the characteristic of these enzymes. These enzymes would be studied in ways that would gauge how successful they would be as an RT in a research or diagnostic environment. These assays would include:

- Researching the enzymes' requirements in terms of reaction buffers, divalent cation usage and optimum reaction temperature
- Assessing the thermostability of the enzyme
- Comparing the ability of the IEP-RTs with the retroviral RTs to see how they perform at reaction temperatures exceeding their optimum
- Measuring the fidelity of the enzyme
- Assaying for additional enzymatic activities such as DNA-dependent DNA polymerase activity
- Measuring the polymerases processivity

- Challenging the polymerases with difficult targets and areas of known secondary structure
- Analysing whether the manipulations made to the enzyme affect any of the above factors in either a negative or positive way.

In addition to these commercially valuable factors it would also be necessary to further the research on the actual proteins. To date no crystallisation of the IEP has been carried out, although their theoretical structure has been compared to that of HIV-1 RT. This research could provide knowledge on the structural basis of these proteins, how they interact with their substrate and how the additional domains beyond the RT have a role in the structure of these proteins.

It was hoped that the IEP would show RT activity at a higher optimum temperature than the wild-type mesophilic RTs due to the thermophilic nature of the organism they would have been isolated from. This could allow the use of a naturally thermophilic enzyme to synthesise cDNA at high temperatures and therefore increase efficiency of reverse transcription.

Chapter 2 – Identification of Thermophilic Bacteria Containing an Intron-Encoded Protein

2.1 - INTRODUCTION

The main RTs available commercially are those from retroviral origin and include MMLV-RT and AMV-RT. These retroviruses infect mesophilic hosts and as a result the enzymes optimum reaction temperature tends to be adapted to the mesophilic temperatures of the hosts at around 37-42ºC. As mentioned previously, these reaction temperatures can cause problems when these RTs are used for cDNA synthesis and complex secondary structures, such as hairpins and stem loops, are found within an RNA template. These RNA structures can cause pausing and dissociation of the enzyme (Harrison *et al.*, 1998; Klasens *et al.*, 1999) leading to a low yield of full-length cDNA and a high yield of truncated products. It has been recognised that there is a need for an RT enzyme that can function at higher temperatures, allowing complex secondary structures to be denatured and higher quality cDNA produced.

Vellore *et al.* (2004) recognised that a potential solution to the problem of mesophilic RTs could be to discover a functioning RT within thermophilic bacteria. The approach they adopted to search for such an RT was to assemble a consensus amino acid sequence from RTs found in a broad range of bacteria which included retrons and Group II Intron ORFs. This consensus sequence was used as the query sequence in a BLAST search, which revealed several previously un-described ORFs with similarities to this sequence. This search identified the *trt* gene, an ORF within a Group II Intron, from the partial genome sequence of *Bacillus stearothermophilus*. The *trt* gene contained the typical RT and maturase domains, as seen in all IEPs, but did not contain the

DNA endonuclease or DNA-binding domains that can be found in other IEPs. Although Vellore *et al*. (2004) successfully cloned the *trt gene*, they found that when they expressed the protein the majority was insoluble and they were only able to partially purify the enzyme. Using a highly-sensitive Product-enhanced RT assay (PERT), Vellore *et al.* (2004) were able to prove that this Trt did indeed have RT activity. This activity was quantified using a second assay that incorporated radiolabelled nucleotides. Using this assay, the enzyme showed maximal RT activity at 65ºC. The activity was however significantly reduced at 75ºC, despite the Trt protein showing detectable RT activity, using the PERT assay, at temperatures as high as 85ºC.

Ng *et al.* (2007) also realised the potential of finding thermostable RTs within thermophilic bacteria and decided to focus on IEPs within Group II Introns. Using the method described below, they identified an IEP within *Bacillus caldolyticus* EA1. Similarly to Vellore *et al.* (2004), this group had difficulties with obtaining soluble protein and achieved a low yield with approximately 10% solubility of this IEP when expressing using a Baculovirus system. Despite this low solubility, they too showed this IEP to have RT activity by the incorporation of radiolabelled nucleotides into poly(rA)●p$(dT)_{12\text{-}18}$ but were unable to detect cDNA synthesis activity, possibly due to the low concentration of enzyme available for the reaction.

The results from these two groups outlined the true potential of finding a thermostable RT within thermophilic bacteria. However, due to the significant level of insoluble protein, both groups concluded that it might be necessary to clone the entire intron element to increase solubility, as other IEPs are known to be more stable when in a complex with their intron RNA. (Zimmerly *et al*., 1999)

In order to identify novel IEPs within thermophilic bacteria for this report, it was decided that the best approach would be to adopt the method used by Ng *et al.* (2007) as this would allow the possibility of finding previously un-sequenced IEPs.

Ng *et al*. (2007) began by making an alignment of mesophilic prokaryotic and eukaryotic IEPs and assessing them for conserved regions across the group. These conserved regions offered potential primer sites for the design of CODEHOP style primers (Rose *et al*., 1998). This style of primer contains a 3' degenerate core that is relatively short with only 3-4 conserved residues required. This allows for codon usage within an organism and, by keeping the degeneracy low, prevents a low yield of individual primers that would otherwise be used up early in the reaction cycles. At the 5' end of these primers is a non-degenerate clamp that allows higher annealing temperatures in later cycles but does not add to the degeneracy of the primer.

Ng *et al.* (2007) designed three CODEHOP style primers, RTF1, RTF2 and RTR1 (Figure 2.1a), based on the alignment of sequences. Two conserved regions were selected for the design of these primers (Figure 2.1b) with the forward primer based on the G(TV)PQG conserved motif and the reverse primer based on the YADD conserved region. Just downstream of the conserved sequence within the forward primer was a second, less conserved motif SPLLANI where the SP was highly conserved. Alignment of the IEP sequences showed these designed primers flanked a region that was highly variable in length, between 130-300bp. A Touch-Down PCR method was used where the high initial annealing temperature was decreased by 1.5ºC per cycle. This cycling protocol increases the specificity of the PCR as only exact matches will anneal to the primers at the higher temperatures. Gradually decreasing this temperature ensures that this specific annealing will take place. When the lower

temperatures are reached, a pool of specific PCR products provide further templates for the subsequent rounds of PCR. Touch-Down PCR products between the correct target sizes were sequenced and analysed for the presence of the conserved SPLLANI motif just downstream of the 5' primer confirming the presence of an RT domain within the bacteria.

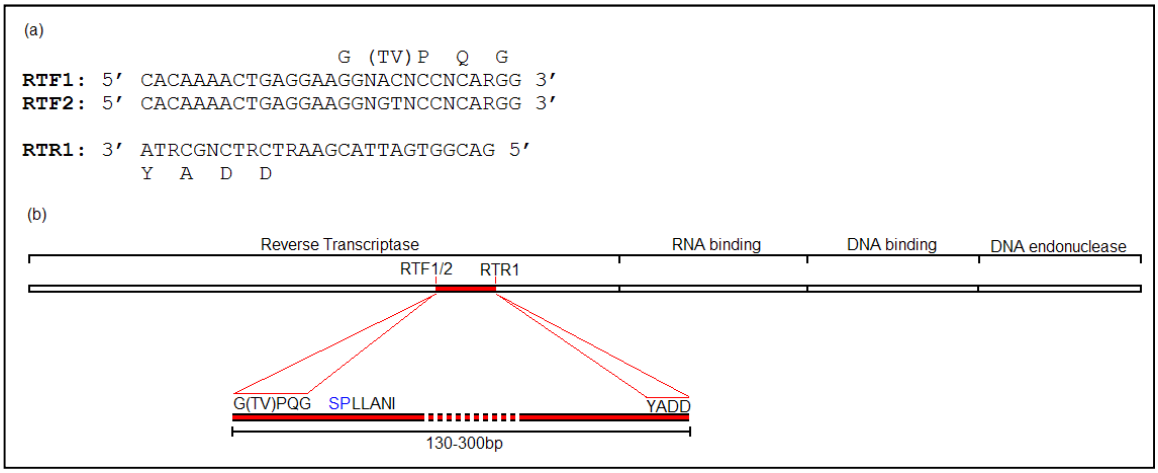

Figure 2.1:

(a) CODEHOP style primers designed by Ng *et al*. (2007) to identify IEP from Group II Introns.

(b) Full-length IEP with all four possible enzymatic domains. The diagram shows the relative position of the primers within the RT. Downstream of RTF1/2 is a second conserved site, SPLLANI, where the SP is very highly conserved and provides confirmation of a reverse transcriptase domain within a Group II Intron-encoded protein.

This diagnostic technique would only identify a small region, 130-300bp, of the IEP gene sequence. It would therefore be necessary to devise a method to discover the remainder of the sequence. The method selected for the purpose of this project was to use an inverse PCR as a Gene-Walking technique (Figure 2.2). PCR normally relies on knowing the 5' and 3' termini of the region to be amplified. However, in this instance, the internal sequence was known with the flanking sequence yet to be identified. This problem was rectified by creating a library of self-ligated gDNA fragments. gDNA was initially fragmented using

restriction endonucleases (RE) and then the digested template circularised in a dilute ligation reaction designed to reduce the formation of multimer, as detailed in the methods for this chapter. Primers were designed using the identified internal fragment facing out toward the unknown sequence. The circularised gDNA therefore provided a template where the primers can direct the DNA polymerase away from the internal fragment toward the unknown flanking sequence. A second-round PCR uses the product from the inverse PCR with a second, nested, set of primers to increase the specificity of the reaction.

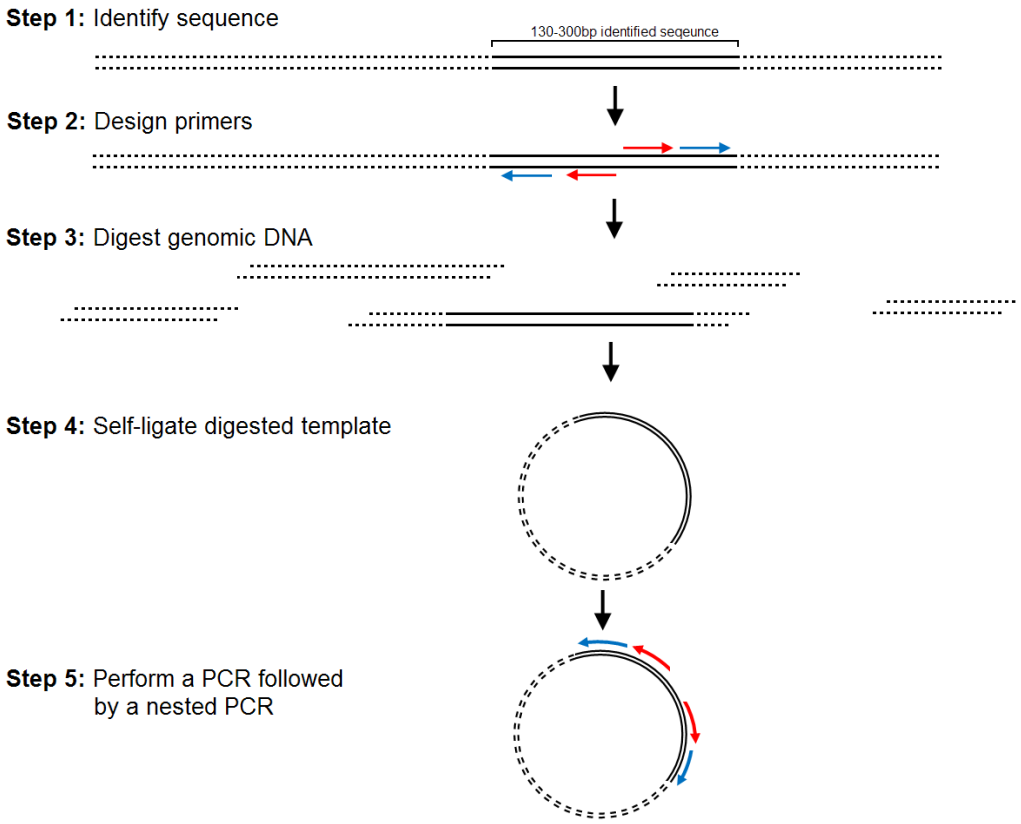

Figure 2.2: A diagrammatic representation of an inverse PCR adopted for a Gene-Walking technique adapted from Triglia *et al*. (1988).

1. The initial 130-300bp of the IEP has to be identified using Touch-Down PCR followed by sequencing.
2. Forward and reverse primers are designed on this identified sequenced facing toward the unknown flanking sequenced.
3. Genomic DNA is digested using restriction enzymes to produce a fragmented library.
4. A dilute ligation reaction is carried out on the digested library to promote circularisation of the template.
5. An inverse PCR is performed using the self-ligated template followed by a second-round PCR using nested primers to increase the specificity of the products.

The product generated from the Gene-Walking PCR could then be TA cloned, as in the methods to this chapter, to allow for sequencing, and then a BLAST search could be used to confirm if the Gene-Walking was successful and additional IEP gene sequence had been obtained. If necessary, Gene-Walking could be repeated with additional primers designed at the ends of the larger

identified fragment. Steps can be made to try to force the direction of the Gene-Walking by designing primers close to the end that is required and by choosing restriction enzymes that cut just before these primer sites so that circularisation would include the majority of the unidentified sequence.

This chapter reports the use of the Touch-Down PCR as a way of identifying IEP genes in thermophilic bacteria, isolated from various soil samples. The inverse PCR Gene-Walking technique would then provide additional sequence information to allow cloning primers to be designed to amplify the IEP gene. Additionally, this chapter also reports on a search for IEP genes within sequenced thermophilic bacteria that would also be studied in this project.

2.2 – MATERIALS AND METHODS

Growth Media

Luria Broth (LB): 10g/litre tryptone, 5g/litre yeast extract, 5g/litre NaCl. With 1.5%(w/v) agar added for LB agar petri dishes. All components were from Melford, Ipswitch, UK.

Tryptone soya broth (TSB): 30g/litre tryptone soya broth (Sigma, Gillingham, UK). 1.5%(w/v) agar was added for TSB agar plates

SOC: 2g/litre Bacto Tryptone, 0.5g/litre Bacto yeast extract (both from BD, Oxford, UK), 10mM NaCl, 2.5mM KCl (Fisher Scientific, Leicestershire, UK). Once autoclaved 10mM $MgCl_2$, 10mM $MgSO_4$ (filter sterilised), and 20mM sterile glucose was added (all Sigma, Gillingham, UK)

YENB: 7.5g/litre Bacto yeast extract, 8g/litre Bacto nutrient broth (BD, Oxford, UK). Medium was made up to correct volume using Milli-Q water.

Buffers

Cell lysis buffer A: 30mM Tris pH8.0 (Fisher Scientific, Leicestershire, UK), 20%(w/v) sucrose (Fisher Scientific, Leicestershire, UK), 1mM EDTA (Fisher Scientific, Leicestershire, UK), 1mg/ml lysozyme (Sigma, Gillingham, UK).

Cell lysis buffer B: 100mM Tris pH8.0, 1%(w/v) Tween20 (Sigma, Gillingham, UK)

Cell lysis buffer C: 1x TE (10mM Tris pH8.0, 1mM EDTA), 7.5%(w/v) Chelex-100 (BioRad, Hemel Hempstead, UK), 0.05M EDTA pH8.0, 0.1%(w/v) SDS (Melford, Ipswitch, UK)].

Sequencing buffer: 400mM Tris-base, 10mM $MgCl_2$

Isolation of Environmental Strains

Soil samples used for the isolation of thermophilic bacterial strains were either collected from hot compost heaps belonging to the Department of Estates, University of Bath, or were previously collected from different international locations at hot temperature sites. 1g of soil was re-suspended in 3ml sterile Milli-Q water, mixed and the contents left to settle for 20min. 100µl of the supernatant was spread onto LB agar plates or TSB agar plates. The agar plates were incubated overnight at various temperatures exceeding 65ºC. Colonies were then re-streaked onto fresh plates and incubated overnight at the original growth temperature. This process was repeated until individual colonies could be obtained.

Environmental Sample Genomic DNA Extraction

In order to release genomic DNA (gDNA) from the bacteria, cell lysis was carried out using a method adapted from Sambrook *et al.* (1989). Individually isolated colonies were re-suspended in 100µl cell lysis buffer A and incubated at 25ºC for 10min. Samples were then microfuged at 13,000xg for 10min and the supernatant discarded. The pellet was re-suspended in 100µl cell lysis buffer B. Samples were heated at 95ºC for 20min and then cooled to -70ºC after briefly mixing. This temperature fluctuation was repeated three times and the samples were then microfuged at 13,000xg for 10min. The supernatant was retained as the source of gDNA. To ensure sufficient cell lysis had occurred, the supernatant was used as a template in a 16S ribosomal RNA (rRNA) gene PCR.

16S Ribosomal RNA Gene Amplification

This PCR used primers based on the 16S rRNA gene, which is highly conserved between different species of bacteria. A positive amplification product was used to indicate that sufficient genomic material had been isolated. The final reaction contained 1x *Taq* mastermix (GeneSys Ltd, Camberley, UK), 25pmol 27F primer and 25pmol 1429R primer (Invitrogen Ltd, Paisley, UK) (Appendix I), 1µl gDNA and sterile Milli-Q water added to a final volume of 50µl. The reaction was cycled under the following conditions:

94ºC	3min	
94ºC	10s	40 cycles
52ºC	10s	
72ºC	2min	
72ºC	7min	

5µl of the reaction was visualized on a 1%(w/v) agarose gel (Lonza Biologics plc, Slough, UK), with the presence of a 1402bp PCR product indicating sufficient gDNA material was present at the start of the reaction.

Genomic DNA Extraction for Lyophilised Samples

This method, taken from Götz *et al*. (2002), was used to extract gDNA from lyophilized bacterial samples obtained from The German Collection of Microorganisms and Cell Cultures (DSMZ). Lyophilised pellets were re-suspended in 567µl of cell lysis buffer C. 100µg proteinase K (Sigma, Gillingham, UK) was added to each sample and then incubated, with occasional mixing, for 60min at 50ºC. Chelex-100 was separated from the supernatant using a microfuge at 1,000xg for 5min and the supernatant transferred to a sterile 1.5ml screw capped microfuge tube. 100µl 5M NaCl and 80µl cetrimonium bromide (CTAB; Sigma, Gillingham, UK) [10%(w/v) in 0.7M NaCl] was added and the sample inverted gently. The samples were incubated for

30min at 65ºC. DNA was then extracted using a phenol/chloroform extraction protocol. An equal volume of buffer-saturated phenol (SureChem Products Ltd, Suffolk, UK) was added to the sample and the mixture vortexed for 1min and then microfuged at 13,000xg for 5min. The upper aqueous phase, containing the DNA, was removed to a fresh microfuge tube and a 1:1 volume of buffer-saturated phenol and chloroform (Surechem Products Ltd, Suffolk, UK) was added. The sample was mixed and microfuged as before. The upper aqueous phase was removed to a fresh microfuge tube and an equal volume of chloroform was added. The samples were mixed and microfuged as before and the chloroform step repeated. The upper aqueous phase was then removed to a fresh tube and an equal volume of isopropanol (Surechem Products Ltd, Suffolk, UK) added, the sample was mixed and immediately microfuged at 13,000xg for 20min. The DNA pellet was washed twice with 70%(v/v) ethanol (Fisher Scientific, Leicestershire, UK), allowed to air dry, and then re-suspended in 20µl TE. An aliquot was electrophoresed on a 1%(w/v) agarose gel to estimate DNA concentration.

If the gDNA yield was low, further gDNA was produced using a GenomiPhi™V2 DNA amplification kit (GE Healthcare, Chalfont St. Giles, UK) as per manufacturing protocol.

Identification of RTs within Un-Sequenced Strains

Extracted gDNA from un-sequenced samples (environmental isolates) was used as a template in a Touch-Down PCR with the following cycling conditions:

94ºC	1min	
94ºC	30s	20 cycles
70ºC (-1.5 ºC/cycle)	30s	
72ºC	30s	
94ºC	30s	30 cycles
55ºC	30s	
72ºC	30s	

This cycling was followed by a final extension step at 72ºC for 20min to ensure sufficient A-tailing by the polymerase. The final reaction contained 1x *Taq:Pfu* (20:1) mastermix (GeneSys Ltd, Camberley, UK), 25pmol RTF1 or RTF2, 25pmol RTR1, 50ng gDNA and water to a final volume of 50µl. Primers, RTF1/2 and RTR1 (Invitrogen Ltd, Paisley, UK), were CODEHOP style primers designed by Ng *et al*. (2007) (Appendix I).

PCR products were analysed on a 2.5%(w/v) agarose gel and any amplicons between 130-300bp were used directly for TA cloning.

TA cloning

3µl PCR product using a *Taq:Pfu* mix of 20:1 was ligated into 0.5µl pCR®2.1 using the TA Cloning® kit (Invitrogen Ltd, Paisley, UK) as per manufacturer's instructions. Ligations were carried out overnight at 16ºC and then heated to 70ºC for 20min to denature the T4 DNA ligase. TA clones could then be ethanol-precipitated before further use.

Ethanol Precipitation of DNA

0.1 volume of 2.5M sodium acetate pH5.2 (Sigma, Gillingham, UK) was added to the sample followed by 2 volumes of 100% ethanol. Samples were incubated at -20ºC for 20-30min and then microfuged at 13,000xg for 10min to pellet the DNA. The DNA pellet was washed with 70%(v/v) ethanol and then microfuged as before. The supernatant was discarded and the DNA pellet left to air dry before re-suspending in nuclease-free water.

Electroporation

Electroporation allowed the transformation of plasmids into various *E. coli* strains.

- pCR®2.1 plasmids were transformed into *E. coli* TOP10F' (New England Biolabs, Hitchin, UK) (Appendix II).
- pET vectors (Novagen® through Merk, Nottingham, UK) were transformed into *E. coli* KRX (Promega, Hampshire, UK) (Appendix II) containing the chloramphenicol-resistant plasmid pRARE2 extracted from Rosetta™2 (Novagen®) (Appendix II). The pRARE2 plasmid supplies rare tRNAs for the codons AUA, AGG, AGA, CUA, CCC, GGA and CGG to reduce protein expression limitations by the codon usage of *E. coli*.
- pUCSacB vector was transformed into TOP10 (NEB Hitchin, UK) (Appendix II).

40µl of electrocompetent *E. coli* cells in 10%(v/v) glycerol (Sigma, Gillingham, UK) were thawed slowly on ice. DNA from ligation reactions was re-suspended in 5µl nuclease-free water before being used in an electroporation 0.5-1µl of ligated, ethanol-precipitated DNA was added to the cells, incubated on ice for 2min and then added to a 0.1mm electroporation cuvette (BioRad, Hemel

Hampstead, UK). Samples were pulsed using the Micropulser™ (BioRad, Hemel Hampstead, UK) at 1.80kV, 3.5-5ms. The cells were then immediately re-suspended in 1ml SOC medium and were incubated at 37ºC with aeration for 1h. Typically, 50-100µl of the 1ml sample was spread onto an LB agar petri dish containing the appropriate antibiotic for the vector used and incubated overnight at 37ºC.

Where TA vector pCR®2.1 was used, samples were spread onto LB agar plates containing 50µg/ml kanamycin, 1mM IPTG, 40µg/ml X-gal (Melford, Ipswitch, UK).

Colonies could then be selected using a sterile toothpick and screened using the colony-screening PCR method, as detailed below, to analyse the cloned DNA fragment. Where pCR®2.1 was used in TA cloning, positive colonies were initially selected using blue/white screening. However, white colonies presumed to be positive were still further screened using the colony-screening PCR.

Preparation of Electrocompetent *E. coli* Cells

10ml LB was inoculated with the appropriate *E. coli* strain and incubated overnight at 37ºC with aeration at 225rpm. 2x 250ml YENB in baffled 500ml shaker flasks were each inoculated with 2.5ml of an overnight *E. coli* culture and incubated at 37°C at 225rpm. Once an OD_{600} of 1 had been reached, the cultures were cooled to 4ºC and the cells pelleted in a pre-chilled centrifuge at 5,000xg for 10min. The supernatant was removed and the cell pellet resuspended in 50ml ice cold sterile Milli-Q water. The cells were centrifuged as before and washed with 50ml ice cold water. The supernatant was discarded and the cell pellet resuspended in 5ml ice cold 10%(v/v) glycerol and centrifuged as before. The supernatant was discarded and the cells

resuspended in 1.5ml ice cold 10%(v/v) glycerol and dispensed into 0.5ml microfuge tubes pre-chilled to -80ºC. The electrocompetent cells were then stored at -80ºC until needed. In order to test the efficiency of the cells, a known concentration of plasmid was electroporated into the cells and a colony count taken of various dilutions of the electroporation sample.

Colony-Screening PCR

Individual colonies were selected with a toothpick and gridded on an appropriate antibiotic LB agar plate, to be retained as a reference plate, and the remaining cells on the toothpick re-suspended in 20µl sterile water. The cells were lysed by heating to 100ºC for 5min and the supernatant used as a template in a screening PCR. Screening primers were dependent on the vector used with M13 forward and reverse primers (Invitrogen LTD, Paisley, UK) (Appendix I) used for pCR®2.1 and T7 promoter and terminator primers (Invitrogen, Paisley, UK) (Appendix I) for pET vectors.

The reactions were set up as follows (final concentrations): 1x *Taq:Pfu* (20:1) mastermix, 25pmol forward and reverse primers, 1µl lysed template and nuclease-free water to a final volume of 50µl. The reactions were cycled as follows

94ºC	3min	
94ºC	10s	} 30 cycles
55ºC	10s	
72ºC	1min/kb	
72ºC	7min	

The PCR product was analysed on an agarose gel to ensure the correct size product was present.

ABI 3100 Sequencing

5-10µl of PCR product to be sequenced was first treated with 1µl ExoSap [5U/µl Exonuclease I (NEB, Hitchin, UK), 0.75U/µl shrimp alkaline phosphatase (Roche, Welwyn Garden City, UK)] and incubated at 37ºC for 15min followed by 80ºC for 15min. The DNA could then be used in a sequencing reaction. Plasmids isolated using a mini-prep kit, in this instance The Wizard® Plus SV mini-prep kit (Promega, Hampshire, UK), could be used directly in the sequencing reaction without the need for an ExoSAP step.

Sequencing reactions were set up as follows: 1µl BigDye® Terminator V3.1 (Applied BioSystems™, Warrington, UK), 7µl sequencing buffer, 3.2pmol sequencing primer (Invitrogen, Paisley, UK), 150-300ng DNA template and nuclease-free water to a final volume of 20µl. Cycle sequencing was carried out with the following conditions:

96ºC	1min	
96ºC	10s	30 cycles
50ºC	5s	
60ºC	4min	
70ºC	20min	

Sequencing reactions were then processed with the addition of 5µl 125mM EDTA and 62µl 100%(v/v) ethanol. Samples were incubated on ice for 30min and then microfuged at 16,000xg for 20min. The supernatant was removed and the DNA pellet washed with 100µl 70%(v/v) ethanol and microfuged as before. The supernatant was removed and the samples air dried. Sequencing reactions were then re-suspended in 10µl Hi-Di™ formamide (Applied BioSystems™, Warrington, UK) and then loaded onto an ABI PRISM® 31000 Genetic Analyzer. Sequencing runs were analysed using Applied BioSystems™ sequence scanner v1.0.

Gene-Walking

Gene-Walking used an inverse PCR method to identify unknown DNA sequence found beyond the boundaries of a known DNA fragment.

Primer Design

Gene-Walking primers were based on known sequence either identified from the initial Touch-Down PCR or from previous rounds of Gene-Walking. Two sets of primers were designed to face out toward the unknown sequence to allow inverse PCR to be carried out (Figure 2.2).

Digest Library

Genomic DNA was digested with selected restriction endonucleases (RE) from New England Biolabs (NEB, Hitchin, UK). RE reactions contained: 1x NEB recommended buffer, 1x BSA where required, 100ng gDNA, 10U RE and made to a final volume of 25µl with nuclease-free water. Reactions were incubated at the optimum enzyme temperature for 3h.

Self Ligation

Digested fragments were ligated in a reaction designed to promote circularisation of fragments instead of multimers. 50ng of the gDNA from each digest was ligated in a reaction containing 1x T4 DNA ligase buffer, 12.5U T4 DNA ligase (NEB, Hitchin, UK) and made up to 100µl with nuclease-free water. Ligations were incubated overnight at 16ºC.

Initial Gene-Walking PCR – Inverse PCR

The self-ligated template was used in a PCR with the inner pair of Gene-Walking primers. The final reaction contained 1x *Taq:Pfu* (20:1) mastermix, 25pmol of forward and reverse primers, 2µl (1ng) of self-ligated gDNA and the final volume made to 50µl with nuclease-free water.

The reactions were cycled using the following conditions:

94ºC	4min		
94ºC	10s	}	35 cycles
55ºC	10s		
72ºC	5min		
72ºC	7min		

Nested Gene-Walking PCR

2µl PCR product from the initial Gene-Walking PCR was used as a template in a nested PCR to increase specificity of the products. The reaction was set up as before but with 30 cycles and a final 72ºC hold for 20min.

TA Cloning

PCR products from the nested PCR were visualised on a 1%(w/v) agarose gel. Any products over 300bp were TA cloned as described previously. Screening and sequencing were carried out as before to identify more of the unknown IEP gene sequence.

Gene-Walking was repeated as necessary to obtain the entire IEP gene sequence

BLAST Searches

Using the NCBI Basic Local Alignment Search Tool (BLAST), sequenced Gene-Walked DNA was used as the query sequence in a search against a database of sequenced bacteria to identify whether it matched IEP gene sequences. The same tool was used to compare known or identified IEPs with others that have been sequenced. The tool was also used to identify sequenced thermophilic bacteria containing an IEP.

2.3 – RESULTS

Non-Sequenced DSMZ Strains

Identification of an IEP

Various, non-sequenced thermophilic bacteria, with growth temperatures ranging from 60-80ºC, were ordered from the DSMZ microorganism collection (Table 2.1).

DSMZ Number	Strain	Growth temperature (ºC)
2178	*Thermodesulfobacterium commune*	70
2913	*Hydrogenobacter hydrogenophilus*	72-76
7242	*Thermolithobacter carboxydivorans*	65
11255	*Carboxydothermus ferrireducens*	65
11347	*Thermodesulfovibrio yellowstonii*	60
11420	*Hydrogenophilus hirschii*	60
11699	*Desulfurobacterium thermolithotrophum*	70
12046	*Hydrogenothermus marinus*	65
12570	*Thermodesulfovibrio islandicus*	65
12571	*Thermodesulfobacterium hveragerdense*	65
14290	*Thermodesulfovacterium hydrogeniphilum*	70
14350	*Persephonella marina*	70
14351	*Persephonella guaymasensis*	70
14484	*Thermocrinus albus*	80
14884	*Marinithermus hydrothermalis*	70
15103	*Persephonella hydorgeniphila*	70
15120	*Sulfurihydrogenibium subterraneum*	62
15241	*Sulfurihydrogenibium azorense*	68
15242	*Caldanaerobacter subterraneus* sub-species *tengcongensis*	75
15286	*Thermodesulfatator indicus*	70
15698	*Thermovibrio ammonificans*	75
16304	*Balnearium lithotrophicum*	70
16510	*Hydrogenivitga caldilitoris*	75
16646	*Thermosediminibacter oceani*	68

Table 2.1: DSMZ bacterial strains and growth temperatures.

The DSMZ bacterial samples arrived lyophilised and the gDNA was extracted using the method taken from Götz *et al*. (2002). Where gDNA concentration was low, the yield was enhanced using the GenomiPhi™V2 DNA amplification kit to ensure enough DNA was present for use in a Touch-Down PCR. The CODEHOP style primers, designed by Ng *et al.* (2007), were used in two individual Touch-Down PCRs. The first reaction contained the primer pair RTF1 with RTR1, and the second contained RTF2 with RTR1 (Appendix I). The products from the Touch-Down PCRs were analysed using a 2.5%(w/v) agarose gel.

0.5µg of 2-log ladder marker (NEB, Hitchin, UK) was used to identify any products between 130-300bp in length which could potentially indicate an RT domain, from a Group II Intron, within the bacteria. Strains 2913, 7242, 11420, 12046, 14484, 14884, 15241, 16510 produced the required product using the primer pair RTF1/RTR2 and strains 15286, 16304 and 16646 produced products within this range using the primer pair RTF2/RTR1. The PCR product from these strains was used directly in a ligation to TA clone the fragments into pCR®2.1. Once ligated, 1µl of the ethanol-precipitated DNA from this reaction was transformed into electrocompetent *E. coli* strain TOP10F'. 100µl of the transformation was plated onto LB agar plates containing 1mM IPTG and 40µg/ml X-Gal allowing for blue-white screening. 10 white colonies from each transformation were selected, gridded and screened using colony-screening PCR with M13 Forward and Reverse primers (Appendix I). The size of the PCR product was analysed on an agarose gel. The PCR products from any successfully screened TA clones were treated with ExoSAP and sequenced following the ABI 3100 sequencing protocol. Despite the fact that several strains were positive for the correct sized product, sequencing of these fragments failed to identify any that contained the conserved motif, SPLLANI, indicative of an RT domain from an Intron-Encoded Protein. Instead, sequencing revealed these products to contain either fragments of genes from

within the organisms, reflecting a lack of specificity during the PCR, or were a result of a multimer formation of the primers.

Non-Sequenced Donated Genomic DNA

Identification of an IEP

GeneSys Ltd (Camberley, UK) donated gDNA from their collection of isolated environmental and DSMZ samples. These samples consisted of unidentified bacteria, presumed to be various *Bacillus* species from Australia and Hong Kong and the known *Bacillus* species, *B. stearothermophilus, B. caldotenax* and *B. caldovelox*. 1ng of these gDNAs was used in the two Touch-Down PCRs as before, and a marker on an agarose gel used to select those strains with potential RT domains (Figure 2.3).

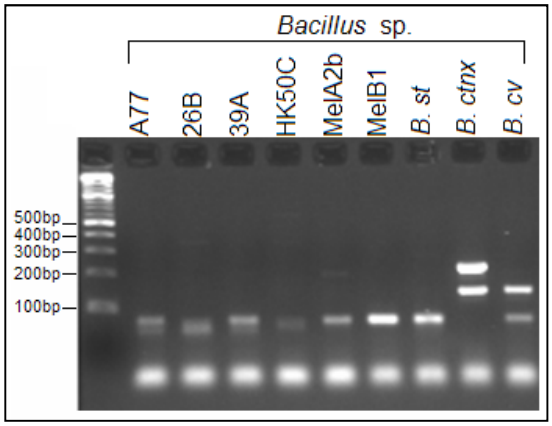

Figure 2.1: A 2.5%(w/v) agarose gel of the Touch-Down PCR products from a thermophilic bacterial screen. The RTF1/RTR1 CODEHOP primer pair positively amplified DNA fragments between 130-300bp in two *Bacillus* species *B. caldotenax* (*B. ctnx*) and *B. caldovelox* (*B. cv*).

The agarose gel (Figure 2.1) showed a background amplification of products of a consistent size within all the reactions that proved negative for the required 130-300bp diagnostic length. The very small, brighter band is consistent with a

primer dimer formation. However; the larger band seen just below the 100bp marker is likely to be due to a multimer formation of the primers as they continually primer against each other and fragments of the template due to their degenerate nature.

B. caldotenax had two products between 130-300bp in length and *B. caldovelox* contained one. In order to see if these were indeed due to amplification of a conserved domain within an IEP, the PCR products were ligated into pCR®2.1. These constructs were transformed into electrocompetent *E. coli* TOP10F' and white colonies from the blue/white screening procedure were selected for further analysis. Once a fragment of the correct size had been confirmed using colony-screening PCR with M13 forward and reverse primers, the PCR product was treated with ExoSAP and used in an ABI3100 sequencing reaction.

B. caldovelox

The sequencing of the Touch-Down PCR identified the following sequence of DNA:

```
cacaaaactgaggaagggacgccacagggagggccgctcagtccactcctgtccaacat
tctcctggatgagctggacaaagaattggaaaaacgagggcacaagtttgtacggtatg
cagacgatttcgtaatcaccgtc
```

Translation of this DNA sequence into a protein sequence (Figure 2.4) revealed similarities to the conserved SPLLANI motif indicating that this strain contained an RT domain from an IEP.

HKTEEGTPQGGPLSPLLSNILLDELDKELEKRGHKFVRYADDFVITV

Figure 2.4: Translated protein sequence from a potential IEP found in *B. caldovelox.* The green highlighted regions indicate the conserved domains from the primer sequence. The yellow highlighted region shows the conserved motif found in RT domains of Group II Introns.

A BLAST search using this acquired protein sequence also verified that the protein matched that of an RT domain within an IEP with the closest match being identical to the *Geobacillus kaustophilus* recombinase of Bh.Int-like element locus tag GK1355-IEP.

This identified internal IEP sequence allowed primers to be designed facing out toward the unknown DNA sequence (Figure 2.5) to allow Gene-Walking to be carried out.

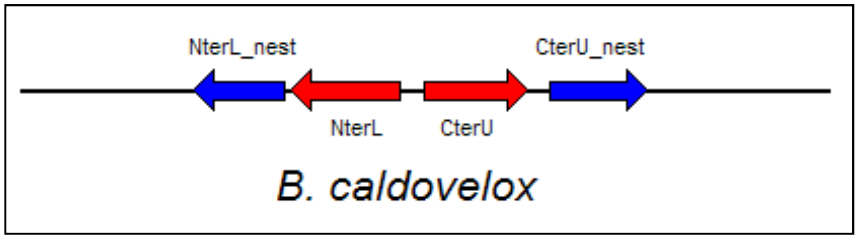

Figure 2.5: A schematic representation of the location of the Gene-Walking primers designed for *B. caldovelox*. Nter and Cter represent the DNA encoding the N-terminus and C-terminus of the protein respectively and U and L relate to upper and lower strand dictating the direction of the primers. Full details of the primers can be found in Appendix I. The red primers are used as a pair in the initial inverse PCR while the blue primers are used as a pair in the second, nested PCR.

B. caldotenax

The Touch-Down PCR fragments from this *B. caldotenax* were TA cloned into pCR®2.1 and sequenced using the ABI 3100 sequencing protocol. Despite

showing two products from the Touch-Down PCR, only one of the products was successfully cloned and identified through sequencing:

```
cacaaaactgaggaaggtacgccacaaggtggcatactctcaccgctgctggccaacat
cgctctgacggtgttggacgcacatttccgtgtgaaatgggatgcccatcggacatctc
agcggcgagatgcccatcgcaaacgtggcggggccacctatcggatcgtccgctacgca
gacgatttcgtaatcaccgtc
```

Translation of this DNA sequence into a protein sequence (Figure 2.6) revealed the conserved motif, SPLLANI, indicative of an RT domain.

```
HKTEEGTPQGGILSPLLANIALTVLDAHFRVKWDAHRTSQRRDAHRKRGGATYRIVRYA
DDFVITV
```

Figure 2.6: Translated protein sequence of a potential IEP found in *B. caldotenax.* The green highlighted regions indicate the conserved domains from the primer sequence. The yellow highlighted region shows the conserved motif found in RT domains of Group II Introns.

This 191bp of identified sequence provided enough sequence information to allow the design of Gene-Walking primers facing out toward the unknown DNA upstream and downstream of this internal fragment (Figure 2.7).

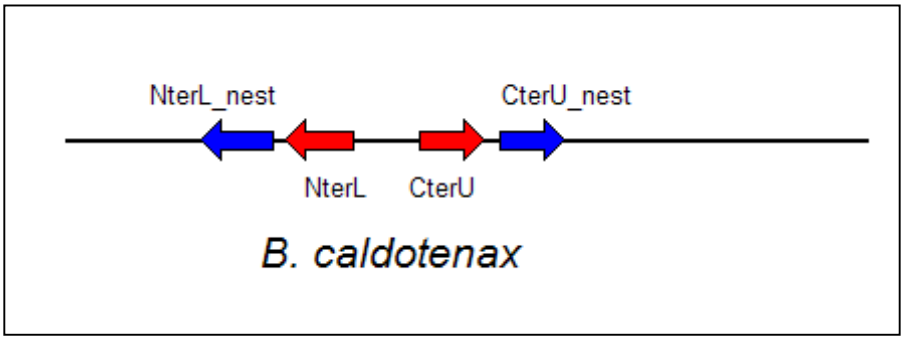

Figure 2.7: A schematic representation of the location of the Gene-Walking primers (Appendix I) designed using the known IEP sequence. Nter and Cter refer to the C-terminus and N-terminus of the protein sequence and the U and L refer to the upper and lower primer sequence, dictating the direction that the primers face. The red primers are used as a pair in the initial inverse PCR while the blue primers are used as a pair in the second, nested PCR.

Gene-Walking

The gDNA from both *Bacillus* species had to be digested with restriction enzymes to produce a fragmented DNA library. 12 REs were initially selected from NEB (Hitchin, UK); these enzymes were chosen because they were 6 cutters and so would not cut the gDNA too frequently, thereby allowing larger fragments to be self-ligated. The enzymes chosen left both sticky and blunt ended fragments. The REs selected for the library were *Kpn* I, *Hin*d III, *Xba* I, *Eco*R I, *Nco* I, *Bam*H I, *Sal* I, *Pst* I, *Eco*R V, *Sac* II, *Aat* II and *Nde* I. Individual digests were carried out on 100ng of the gDNA and the template was digested for 3h to ensure complete digestion. A dilute ligation, designed to promote self-ligation and reduce the likelihood of multimers forming, was set up with 50ng of the digested gDNA. This ligation could then be used as a template in the initial Gene-Walking PCR.

Gene-Walking was carried out using the Gene-Walking primers, designed to face upstream and downstream of the known sequence, and the self-ligated gDNA library. Two rounds of PCRs were required, the first using the inner primer set in an inverse PCR and the second using template from the first PCR with the nested primer pair. This second PCR allowed an enhancement of the specificity of the Gene-Walking, reducing the yield of non-specific fragments. Any fragments greater than 300bp from the second Gene-Walking PCR were cloned into the TA vector, pCR®2.1, transformed into TOP10F' and screened using colony-screening PCR. Any fragments of suitable size were then sequenced using the ABI 3100 sequencing protocol to identify possible IEP gene sequences.

Gene-Walking of *B. caldotenax* gave several bands greater than 300bp in length from the digest libraries of *Bam*H I, *Eco*R I and *Hin*d III. However, sequencing of these fragments and using them as the query sequence in a BLAST search

showed that, although the fragments matched other *Bacillus* genes, none of them matched any known IEP within the search database. Increasing the annealing temperature of the primers in the Gene-Walking PCR to increase specificity resulted in no products greater than 300bp.

Gene-Walking of the *B. caldovelox* strain gave several bands greater than 300bp from digest libraries *Aat* II, *Kpn* I, *Nde* I and *Xba* I. These bands were TA cloned into the pCR®2.1 vector, transformed into TOP10F' and sequenced to identify the fragments. All fragments from these digests corresponded to an IEP. Piecing together the different fragments revealed a total of 642bp of IEP gene sequence (Appendix III). When using the NCBI BLAST tool, this 642bp of acquired sequence matched identically to an IEP within *Geobacillus kaustophilus* HTA426, the recombinase of Bh.Int-like element (Appendix III), locus tag GK1355. From here on this *G. kaustophilus* IEP will be referred to as the GK1355-IEP.

Non-Sequenced Environmental Strains

For this investigation into novel IEPs, it was necessary to isolate different thermophilic bacteria in the hope they would contain a Group II Intron with an IEP. Initially, samples from hot compost were collected from the University of Bath with permission from the Department of Estates. The samples from this collection ranged from the oldest, at approximately 3 years old, to the youngest of approximately 6 months old. The hottest samples were retrieved from regions estimated to be 1 year old. 1g of material from these samples was resuspended in water and the supernatant spread on both LB and TSB agar petri dishes. These petri dishes were incubated overnight at temperatures of 65, 70 and 75ºC. No thermophilic bacteria were isolated at the 70 and 75ºC temperature ranges. Therefore, all thermophilic bacteria isolated and studied were incubated at 65ºC. The initial plating of the supernatant from the 65ºC

incubation produced a lawn of mixed bacterial types. These were re-plated onto fresh agar until individual colonies could be isolated. Individual colonies, with differing morphologies, were then lysed to extract the gDNA, and the supernatant tested using a 16S rRNA PCR, to ensure that enough gDNA was present for amplification. In total, 20 different bacterial strains, four from each age of compost, were successfully lysed and their gDNA used as template in two Touch-Down PCRs, one with RTF1/RTR1 primers and the other with RTF2/RTR1 primers. The samples were analysed on an agarose gel and comparison with markers revealed none of the samples to contain a fragment between 130-300bp in length, suggesting none of the isolated strains from this compost contained an IEP with an RT domain.

In order to expand the search for thermophilic bacteria containing an IEP, additional soil samples had to be screened. The additional samples were from the collection belonging to the Centre for Extremophile Research, University of Bath and were stored at 4ºC. The soil samples that were used were from Lanzarote, Melbourne Yarrow River, Kuala Lumpur, New Zealand and Bali. A total of 16 individual strains were isolated from these samples, all growing on either LB or TSB plates at 65ºC. The gDNA from these strains was used in two Touch-Down PCRs with the two different primer sets to identify any strains with an RT domain (Figure 2.8).

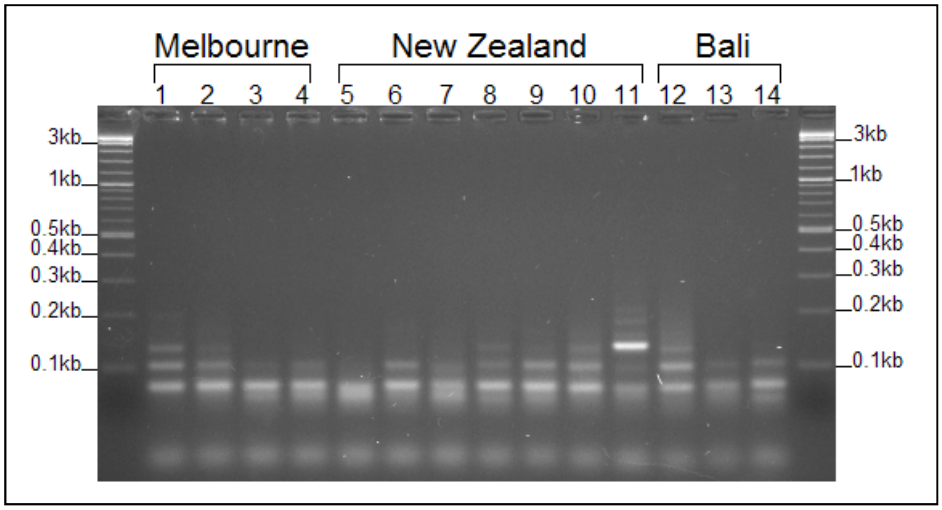

Figure 2.8: An agarose gel of the products from a Touch-Down PCR and RTF2/RTR1 primer pair. Isolate numbers 1 and 2 from Melbourne, and 8,9,10 and 11 from New Zealand have a product between 130-300bp in length.

The Touch-Down PCR products from isolates 1 and 2 (from Melbourne) and 8, 9, 10 and 11 (from New Zealand), all at the critical size between 130-300bp, were cloned into the TA vector pCR®2.1, transformed into TOP10F' and then sequenced. Out of the strains sequenced, New Zealand strain 11 which grew on TSB was shown to contain a potential IEP with an RT domain.

New Zealand Strain 11

The cloned fragment from the Touch-Down PCR was sequenced and revealed the following DNA sequence:

cctcaaggcggtcccctcagcccctgctggcgaacatccttctcgacgatttagacaa
ggagttggagaagcgtggattgaaattctgccgctacgctgacgat

Translation of this DNA sequence into a protein sequence (Figure 2.9) revealed the conserved SPLLANI motif indicative of an RT domain.

PQGGPLSPLLANILLDDLDKELEKRGLKFCRYADD

Figure 2.9: Translated protein sequence of a potential IEP found in a thermophilic bacterial strain isolated from New Zealand. The green highlighted regions indicate the conserved domains from the primer sequence. The yellow highlighted region shows the conserved motif found in RT domains of Group II Introns.

16S rRNA sequencing of this isolate from New Zealand showed it to be a strain of *B. stearothermophilus*. The identification of this species was reflected in the fact that a BLAST search of this obtained IEP protein sequence matched identically with that from the sequenced *B. stearothermophilus Ttr* gene that is an Intron-Encoded Protein with an RT domain. However, on a DNA sequence level, there were differences between the two strains. Using the sequence obtained from the Touch-Down PCR, primers were designed to face out toward the unknown IEP gene sequence (Figure 2.10).

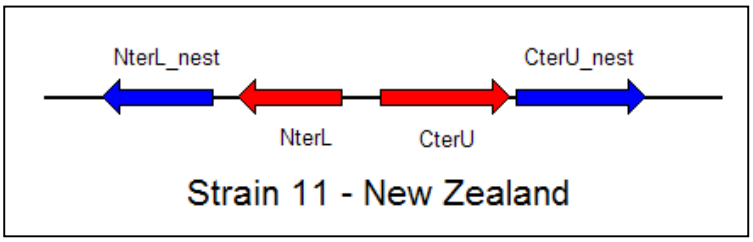

Figure 2.10: A schematic of the location of the Gene-Walking primers (Appendix I) designed using the known IEP sequence. Nter and Cter refer to the C-terminus and N-terminus of the protein sequence and the U and L refer to upper and lower primer sequence, dictating the direction that the primers face. The red primers are used as a pair in the initial PCR while the blue primers are used as a pair in the second, nested PCR.

In order to create a library of digested gDNA, it was necessary to obtain a purer and higher yield of DNA than that provided by the cell lysis step. Initially, 10ml of TSB were inoculated with a single colony of strain 11 and incubated overnight with aeration at 65ºC. This strain failed to grow under these conditions;

however, incubation of the culture at 65ºC overnight with no aeration allowed the strain to grow in the TSB medium. Once the cells had been pelleted by centrifugation, the DNA was extracted using the DNeasy Tissue Kit (Qiagen, Crawley, UK) following the protocol for the isolation of genomic DNA from Gram-positive bacteria. 12 REs were selected from NEB to digest the gDNA and create a fragmented gDNA library. REs that mainly used 6 bases as a recognition sequence were selected, as before, to prevent the gDNA becoming too fragmented and thereby to allow fragments greater than 300bp to be cloned. The enzymes used for the digest library were *Aat* II, *Bam*H I, *Eco*R I, *Hin*D III, *Kpn* I, *Msp* I (4 base recognition sequence), *Nco I*, *Nde* I, *Nhe* I, *Nsi* I, *Sal* I and *Pst* I. 100ng of the gDNA was digested with each enzyme in 12 individual digests. The restriction enzymes were then heat-killed according to NEB guidelines and 50ng of the digested gDNA used in a dilute ligation reaction designed to promote circularisation of the fragments. The Gene-Walking primers designed using the known IEP gene sequence were then used in two Gene-Walking PCRs. The second PCR used the product from the first PCR and the nested primers allowed an increase in specificity. The product from the second-round Gene-Walking was analysed on an agarose gel (Figure 2.11). Products from the digest library from *Aat* II, *Eco*R I, *Nco* I and *Sal* I were above 300bp in length and were cloned into TA vector pCR®2.1, transformed into TOP10F' and screened using blue-white screening and colony PCR.

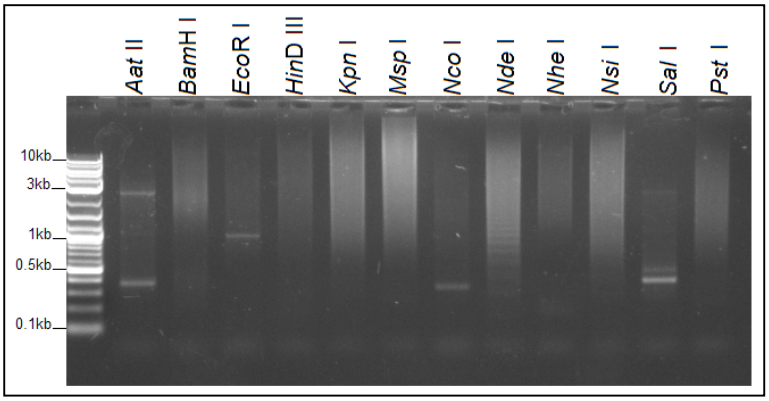

Figure 2.11: A 2.5%(w/v) agarose gel of the second-round Gene-Walking PCR. Fragments from digests *Aat* II, *Eco*R I, *Nco I* and *Sal* I were used in ligation reaction as their products were greater than 400bp in length.

The agarose gel from the second round gene walking sometimes included multiple fragments for each digested library. This is likely to be due to two separate aspects of the process. Firstly incomplete digestion of the genomic template could result in more than one different sized fragments being produced with the same primer set. Alternatively, the ligation reaction itself could result in some multimer formation, as well as circularisation of the fragments, therefore producing additional size products with the same primer sites. The smears can be seen in some of the reactions were no specific product is produced could be down to the presence of undigested gDNA.

The sequencing results of all these fragments gave sections of the IEP within this New Zealand strain. Piecing together these fragments gave 936bp of the IEP gene sequence from this *Bacillus*. When using the ClustalX multiple sequence alignment programme, this sequence was aligned against the *B. stearothermophilus Trt* gene and was shown to contain the ATG start site. However, the sequence obtained so far did not reach the 3' end and therefore the stop codon for this gene. Using the NCBI nucleotide BLAST program, the gene sequence from the New Zealand strain was shown to have 97 % identity

to the *Trt* gene. New Gene-Walking primers, 7.6_GW2, were designed based on the new sequencing information from the 3' end of the gene to try to identify the 3' end, downstream from that already identified, and to reach the stop codon (Figure 2.12).

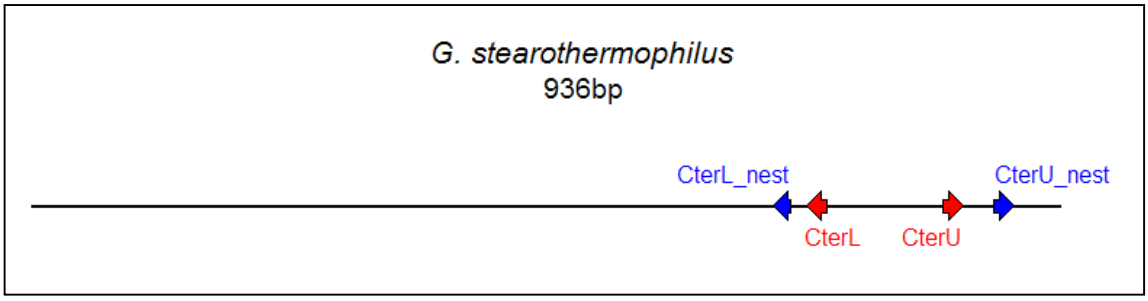

Figure 2.12: A schematic representation showing the location of the second set of Gene-Walking primers for the New Zealand strain of *B. stearothermophilus.* The primers were designed at the 3' end of the known gene sequence to allow Gene-Walking into the remainder of the IEP gene sequence. Cter refers to the C-terminus of the protein and U and L denotes the upper and lower DNA strand respectively and therefore the direction of the primers.

The digested, self-ligated libraries of gDNA were again used as template in a Gene-Walking PCR with the new set of primers. However, Gene-Walking proved to be unsuccessful with the new primers being non-specific to the IEP gene sequence. New Gene-Walking primers, 7.6_GW3 (Appendix I) were designed to try to increase the specificity of the Gene-Walking reactions; however, despite the fact that some of the REs were known to cut just upstream of the primer locations, these new primer sets failed to obtain any new IEP gene sequence.

Sequenced DSMZ strains

Investigations into sequenced thermophilic bacteria containing IEPs were also carried out. Using the NCBI protein BLAST tool, the *G. kaustophilus* HTA426

GK1355-IEP element was compared to other known IEPs within the database. The majority of the matches were from mesophilic organisms with growth temperatures ranging from 28-37ºC. However, the BLAST search did reveal two thermophilic bacteria with potentially thermostable IEPs:

- *Thermosinus carboxydivorans* Nor1 had an optimum growth temperature of 60ºC and contained one IEP gene sequence with 63% protein identity to the *G. kaustophilus* HTA426 GK1355-IEP.
- *Petrotoga mobilis* had an optimum growth temperature of 55ºC and contained 3 IEPs within its genome, two with a 54% identity and one with a 53% identity to the *G. kaustophilus* HTA426 GK1355-IEP.

It was hoped that these IEP gene sequences would code for an IEP capable of showing RT activity at high temperatures.

2.4 – DISCUSSION

In this chapter, a number of methods were described to identify thermophilic bacteria containing an IEP with an RT domain. Initially, un-sequenced bacteria were screened for a conserved motif found within the RT domain. This screening method was based on the same method used by Ng *et al*. (2007) where CODEHOP primers, based on alignments of bacterial RTs, were designed to allow the amplification of a conserved region within the RT. When Ng *et al*. (2007) used this technique the group identified 6 thermophilic bacteria, out of the 34 studied, that contained this RT motif. The thermophilic bacteria containing IEPs identified by this group included *Bacillus caldolyticus EA1; Bacillus caldolyticus* YP-T; *Caldibacillus cellulovorans* (CompA2)*; Thermoanaerobium* species TOK 6B.1*; Clostridium thermohydrosulfurium* species RT66B.1; and *Thermus thermophilus* HB8*.* When this method was adopted for this project, the occurrence of these IEPs was found to be lower than that suggested by Ng *et al*. (2007).

An initial screening of 24 thermophilic bacteria from the DSMZ revealed that none of them contained this conserved domain. The search for an IEP within thermophilic bacteria was then expanded to include additional gDNA from thermophiles isolated by GeneSys Ltd. This search revealed two *Bacillus* species to contain a possible IEP. *B. caldotenax* contained two possible RTs, of which the conserved region from one of these IEP was identified from sequencing, and *B. caldovelox* contained one possible IEP.

In order to expand the search further, unidentified thermophilic bacteria were isolated from a hot compost heap belonging to the Department of Estates, University of Bath as well as from soil samples belonging to the Centre for

Extremophile Research, University of Bath. The samples from the collection from the Centre for Extremophile Research were from Lanzarote, Melbourne Yarrow River, Kuala Lumpur, New Zealand and Bali. In all cases it proved difficult to isolate any bacteria growing at temperatures above 65ºC. It is possible that any bacteria that grow at temperatures exceeding the 65ºC would require different media with additional supplements, such as trace elements or minerals, which were not provided in the media used in this study. Altering the media could allow the broadening of the search for thermophilic bacteria with an IEP. In total 36 different thermophilic bacteria, with growth temperatures of 65ºC, were isolated and screened using Touch-Down PCR as a method of identifying any bacteria containing an IEP. Of the 36 isolates, one strain from New Zealand, a *B. stearothermophilus* strain, was revealed to contain the conserved region indicative of an RT domain from an IEP.

Once the conserved regions were identified, Gene-Walking primers were designed to allow an inverse PCR to be carried out to discover the remaining IEP gene sequence. These Gene-Walking primers were designed against one of the RT fragments from *B. caldotenax*, the fragment discovered in *B. caldovelox* and the fragment identified in the New Zealand strain 11, *B. stearothermophilus*. Once the gDNA of these strains had been fragmented by RE digest and circularised in a ligation reaction, inverse PCR was carried out followed by a nested PCR to increase specificity in the identification of additional IEP gene sequence.

B. caldotenax

Large fragments of DNA were isolated from the two rounds of Gene-Walking PCR. However, cloning revealed that, although they matched *Bacillus* genes, they were not specific to the IEP. Attempts were made to improve the specificity of the Gene-Walking including increasing the annealing temperature of the

primers. However, no new IEP gene sequence, beyond the boundaries of the conserved fragment, could be identified. Due to the limitations of size of this conserved fragment, it was not possible to re-design new Gene-Walking primers and as a result, this IEP was left unidentified.

B. caldovelox

The Gene-Walking PCRs from the conserved IEP gene fragment gave several large PCR products to be sequenced. Several of these products revealed new IEP gene sequence and when pieced together, this technique allowed an identification of 642 bases of IEP gene sequence. Using the BLAST tool revealed them not only to match an IEP, but also to be identical to an IEP within *G. kaustophilus* HTA426. This gene, a recombinase of Bh.Int-like element referred to as GK1355-IEP, included the RT domain and maturase domain typical of that seen in IEP and which had not been previously studied. Using the knowledge that the IEP gene sequence obtained so far from *B. caldovelox* was identical to this fully-sequenced gene, no further Gene-Walking was necessary.

Bacillus stearothermophilus, New Zealand strain 11

The first round of Gene-Walking of this strain proved to be very successful with several large DNA fragments obtained. Sequencing and piecing together these fragments revealed 936bp of IEP gene sequence which, when translated into a protein sequence, had a 97% identity to the known Trt protein from *B. stearothermophilus*. These 936bp identified included the start codon of the gene; however, the stop codon had not yet been reached. Attempts were made to Gene-Walk to the 3' end of the gene. This included designing primers specifically at this end of the gene, ensuring that the gDNA digest library included enzymes that cut just upstream of these primer sites and two different primer sets being designed for Gene-Walking. However, despite these efforts,

the 3' of this gene was not fully identified. Any cloning primers would therefore have to be based on the known sequence of the *trt* gene.

The work presented in this chapter revealed a number of thermophilic bacteria that contain an IEP. Screening of un-sequenced isolates found three potential IEP with RT activity. *B. caldotenax* contained an IEP that was not successfully Gene-Walked. However, *B. caldovelox* and a New Zealand strain of *B. stearothermophilus* were Gene-Walked successfully and the identification of these genes would allow their cloning and subsequent expression of the IEP. Using a BLAST search also revealed two additional sequenced thermophilic bacteria that contain IEPs. *P. mobilis* contained three IEPs and *T. carboxydivorans* contained one. The sequence of these would allow the design of primers to allow the cloning of these genes and the expression of the proteins. It was hoped that with a variety of different IEPs identified they would show RT activity at the temperatures at which the bacteria grow and therefore have a potential use as a thermophilic reverse transcriptase.

Chapter 3 – Cloning of Intron-Encoded Protein Genes, Manipulation and Protein Expression

3.1 – INTRODUCTION

Intron-Encoded Proteins in Thermophilic Organisms

The screening of unidentified or un-sequenced bacteria carried out in chapter 2 revealed IEPs from different *Bacillus* and *Geobacillus* species. These genera are extensive, containing rod-shaped bacteria with a diverse range of growth conditions and capable of forming endospores. These can include aerobic to faculatively anaerobic organisms and growth conditions can range through the extremes of temperature, pH and salt concentration (Nazina *et al.* 2001). Two *Bacillus* species, *B. caldovelox* and *B. caldotenax*, discovered in superheated pool water in the USA by W. Heinen (DSMZ online), were identified as containing IEPs. The partial IEP gene sequence identified within *B. caldovelox* was identical to the GK1355-IEP from *G. kaustophilus* HTA426. This was surprising due to their geographical separation as the *G. kaustophilus* strain HTA426 strain was isolated from deep sea mud of the Mariana Trench (Takami *et al*. 2004). The GK1355-IEP contains the typical RT and maturase domains; however, it lacks the DNA-binding and endonuclease domains at the C-terminus.

Other *Bacillus/Geobacillus* species known to contain IEPs include the *B. caldolyticus* EA1 IEP (mentioned in chapter 2), which is almost identical to another, *G. kaustophilus* IEP, GK1296-IEP. These IEPs differ by two nucleotides, causing one amino acid alteration, and an apparent frame-shift at

amino acid 38 in GK1296-IEP. At present, it is unclear whether this frame-shift is genuine or due to a sequencing error (Ng *et al.* 2004). The GK.Int 1 intron, containing the GK1296-IEP, was found to interrupt a *recA* gene (Chee and Takami. 2005) (Figure 3.1) and EA1 IEP also interrupts a gene with high homology to *recA.* RecA is considered ubiquitous and multifunctional with involvement in homologous recombination, DNA repair and the SOS response. GK.Int 1 intron was found to self-splice from the *recA* transcript, restoring its activity (Chee and Takami. 2005). Both the EA1 and the GK1296-IEP contain the RT and maturase domains, lack a DNA-binding domain but contain an endonuclease domain at the C-terminus of the protein.

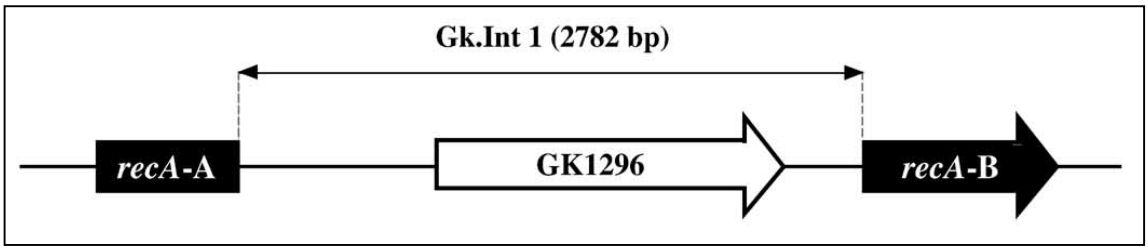

Figure 3.1: A schematic diagram of the *recA* gene of *G. kaustophilus* interrupted by a 2782bp self-splicing Group II Intron (Gk.Int 1). The arrow marked GK1296 corresponds to the IEP within the Group II Intron. (Chee and Takami. 2005).

The BLAST search detailed in chapter 2 identified an IEP within the thermophile *T. carboxydivorans*. This thermophile has an optimum growth temperature of 60ºC and was isolated from a small mud and water pool in the Norris Basin of Yellowstone National Park (Sololova *et al.* 2004). This bacterium is a curved motile rod with a cell wall like that of Gram-negative bacteria; however, its 16S rRNA sequence revealed it to comprise a new genus, within the *Bacillus-Clostridium* phylum of the Gram-positive bacteria (Sokolova *et al.* 2004). The IEP within this bacterial strain contained the typical RT and maturase domains seen in all bacterial group II IEPs. However, the DNA-binding and endonuclease domains were absent.

The BLAST search also revealed the presence of three IEPs within the thermophile *Petotoga mobilis*. This thermophile, with an optimum growth temperature between 58-60ºC, was isolated from hot oil-field water of a North Sea oil reservoir (Lien *et al.* 1998). This bacterium is a rod-shaped cell with a characteristic sheath-like outer cell membrane ('toga'), typical of the *Thermotoga* species. However, unlike other members of the order *Thermotogales*, *P. mobilis* is able to tolerate higher salt concentrations and 16S rRNA sequencing revealed it to be a distinct lineage within this order (Lien *et al.* 1998). The three IEPs within this bacterium show high similarity to each other and although they contain the typical RT and maturase domains, they do not contain the DNA-binding or endonuclease domains.

It was hoped that, with the variety of different IEPs discovered, some would show potential as an active soluble protein with an optimum RT activity reflecting the growth temperatures of the organism from which they originate.

Protein Engineering

Protein engineering has been performed on DNA polymerases as a way of improving their commercial value and broadening their usage. These approaches can include increasing the resistance of a polymerase to inhibitors that could appear in an impure DNA sample to coupling characteristics such as high-fidelity with high processivity, two desirable features that do not often naturally occur together. The latter enhancement would aim to use an enzyme with known high-fidelity, such as *Pfu* DNA polymerase, and increase its processivity to prevent limitations with the size of fragment that can be amplified. Enhancing the processivity of a polymerase will mean that more nucleotides are incorporated during a single enzymatic event before the dissociation of the enzyme (Davidson *et al*. 2003) and has the potential of enhancing the length of amplification product that can be achieved as well as an

additional affect of increasing the speed of the reaction. Possible solutions have included:

- Addition of a hairpin-helix-hairpin (H-h-H) motif – this sequence is found in a variety of DNA-binding proteins. As the motif contacts the sugar-phosphate backbone of the DNA rather than individual bases (Pavlov *et al.* 2002), these proteins are able to bind DNA in a non-specific manner. However, the addition of H-h-H domains within DNA polymerases has produced mixed results and although processivity in *Pfu* was enhanced, no improvement was seen with *Taq* DNA polymerase.
- Insertion of a thioredoxin binding domain – with no natural affinity to either ss or dsDNA (Huber *et al*. 1986), thioredoxin is thought to act as a processivity factor, clamping the polymerase onto the DNA (Bedford *et al.* 1997). T7 DNA polymerase binding *in vitro* to added *E. coli* thioredoxin will have enhanced processivity from usually no more than 12 to over thousands of nucleotides (Tabor *et al.* 1987). However, introduction of a thioredoxin binding domain into an analogous position within *Taq* DNA polymerase greatly lessened the activity of the enzyme (Davidson *et al* 2003) and therefore its use in PCR is yet to be proven
- Insertion of a dsDNA-binding domain – Sso7d is a small dsDNA-binding protein found in *Sulfolobus solfataricus*. Its role in Nature is believed to mimic that of eukaryotic histones (McAfee *et al*. 1995), stabilising the DNA of the thermophilic organism at their high growth temperature. Sso7d fused onto the C-terminus of *Taq* or the N-terminus of *Pfu* DNA polymerase has been shown to greatly increase the processivity of these polymerases (Wang *et al*. 2004).

Mesophilic RTs have also been targeted for modifications to improve their performance in synthesising cDNA. Mutations within RTs tend to be made to improve the efficiency of cDNA production. Common manipulations were discussed in detail in chapter 1 and include:

- RNase H minus mutants – RTs from retroviruses contain an additional RNase H domain that functions to degrade the RNA strand of an RNA:DNA hybrid. This activity, when present in first-strand synthesis reactions, prevents high yields of full-length cDNA and can eliminate the potential of repeated copying of the RNA template (Berger *et al*. 1983). This RNase H activity can both increase the level of RNA required as a template and can lead to this template becoming fragmented.
- Thermostability mutations – running reactions at a higher temperature allows the melting of RNA secondary structure that can act to stall the enzyme leading to its dissociation from the template. A variety of mutations are found in commercially-available RTs, such as that in SuperScript III (Invitrogen) where the enzyme's optimum reaction temperature has been increased to 50ºC.

It was hypothesised that the IEPs discovered in this report would be more thermostable and thermoactive than the mesophilic retroviral RTs and that the absence of an RNase H domain could further reduce the problem of truncated cDNA synthesis. However, a method to increase processivity by reducing the occurrence of dissociation of the RT from the template, therefore increasing the amount of nucleotides incorporated during one enzymatic event, was deemed as being desirable.

A single-stranded binding protein (SSB) from *Thermus thermophilus, Tth*SSB, was shown to have an unexpected influence on the efficiency of cDNA synthesis by the *T. thermophilus* DNA polymerases. In the absence of *Tth*SSB, *T. thermophilus* DNA polymerase produced cDNA fragments of at least 264bp. However, this polymerase failed to produce cDNA where a larger 1085bp target was used. With the addition of *Tth*SSB to the cDNA synthesis reaction DNA products up to 3285bp could be detected by PCR (Perales *et al*. 2003). Perales

et al. (2003) hypothesised that this would be due to the *Tth*SSB having an unfolding effect on the RNA template. It was therefore agreed that a DNA-binding protein could serve to enhance the processivity of an RT enzyme. dsDNA-binding proteins from *Sulfolobus* species were selected as ideal candidates due to their high thermostability and their proven use with DNA polymerases in PCR (Wang *et al*. 2004). Although Wang *et al.* (2004) used the dsDNA-binding protein Sso7d to improve the processivity of polymerases, Sac7d from *S. solfataricus* was selected for this project as, unlike Sso7d, it has not been associated with ribonuclease activity (Fusi *et al*. 1993).

A PCR approach would have to be adopted to allow the artificial fusion of two different genes into one ORF. The technique that was used for this report was an overlap extension PCR (reviewed by Ling and Robinson, 1997) (Figure 3.2).

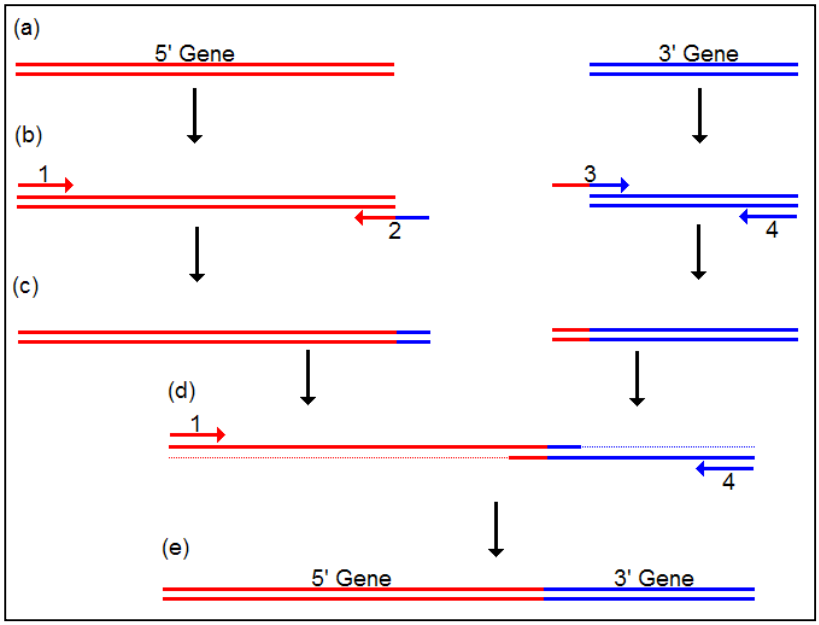

Figure 3.2: A diagram showing an overlap extension PCR method

(a) The sequences of the two genes, one at the 5' end and one at the 3' end of the final fused product, have to be known

(b) Four primers are designed, two for each gene.

Primer 1: Contains the start codon and downstream sequence of the 5' gene.

Primer 2: Overlaps the two genes with the primer including the end of the 5' gene (stop codon removed) extending beyond the start of the 3' gene.

Primer 3: An antisense version of primer 2.

Primer 4: Contains gene specific sequence of the 3' gene followed by the stop codon.

Two PCRs are performed. The first uses the 5' gene template with primer pair 1 and 2 and the second uses the 3' gene template with primer pair 3 and 4.

(c) Two PCR amplicons produced:

1. 5' gene with a removed stop codon and 20-25 bases of the start of the 3' gene.
2. 3' gene with an upstream region containing 20-25 bases of the end of the 5' gene.

(d) The products from the first two PCRs are mixed together and used as the template in the final PCR, to fuse the two genes together, using primers 1 and 4 for amplification.

(e) The final product fuses two domains together in one ORF allowing the production of a single protein.

This technique involves the amplification of individual genes in separate PCRs with primers adding extended bases matching the region to be fused. The annealing step in the final PCR allows the overlapping regions to hybridise and the extension step allows the DNA polymerase to extend along the remainder of the gene. The resulting product is two fused genes in one ORF that will be translated as a single protein.

3.2 – MATERIALS AND METHODS

Buffers

Cell lysis buffer D: 50mM Tris-HCl, pH8.0, 100mM NaCl, 1mM EDTA

Protein loading buffer: 0.1M Tris pH6.8, 20%(w/v) sucrose 4%(w/v), SDS, 0.2%(w/v) bromophenol blue (Sigma, Gillingham, UK)]

SDS PAGE gel running buffer: 52mM Tris-base, 0.1%(w/v) SDS, 0.4%(w/v) glycine

SDS PAGE gel stain: 0.025%(w/v) Coomasie brilliant blue (R250) (Sigma, Gillingham, UK), 45%(v/v) methanol, 10%(v/v) glacial acetic acid (Fisher Scientific, Leicestershire, UK).

SDS PAGE gel de-stain: 20%(v/v) methanol, 10%(v/v) glacial acetic acid

Cloning Primer Design

Cloning primers were designed based on a known gene sequence to allow amplification of the gene and subsequent ligation into an appropriate vector.

Forward Cloning Primers

These contained six random nucleotides (N) followed by an *Nde* I site making use of the natural ATG start site in the gene, and then 18-25 bases of gene specific sequence directly downstream of the ATG start site.

5'- (N_6) CAT ATG (gene specific nucleotides$_{18-25}$) - 3'

Non His-tagged Reverse Primer

These contained 18-25 bases of gene specific sequence ending in the gene's natural stop codon (altered to TAA if necessary), followed by an *Xho* I site and six additional random nucleotides (N) to allow cleavage with the *Xho* I RE.

3' - (gene specific nucleotides$_{18\text{-}25}$) TAA CTC GAG (N_6) - 3'

Reverse Primer to Allow C-Terminal His-Tagging

These contained 18-25 bases of gene specific sequence, and the gene's natural stop codon was deleted and replaced with an *Xho* I site. This would utilise the 6his-tag found in pET24a(+) when using the *Xho* I site at the 3' end of the gene.

3' - (gene specific nucleotides$_{18\text{-}25}$) [TAA replaced with CTC] GAG (N_6) - 3'

Gene Amplification

Phusion® DNA polymerase (NEB, Hitchin, UK) was the enzyme of choice for gene amplification due to its high-fidelity, with an error rate 50-fold lower than that of *Taq*. Components of the PCR included 1x Phusion® High-Fidelity (HF) buffer, 0.2mM dNTPs (GeneSys Ltd, Camberley, Surrey), 50pmol upper and lower primers, 50ng gDNA, and 2U Phusion® DNA polymerase. The final volume was made up to 100µl with nuclease-free water. The reactions were cycled as follows

98ºC	30s	
98ºC	10s	} 20 cycles
Primer Tm+3ºC	10s	
72ºC	30s/kb	
72ºC	7min	

PCR products were visualised on a 1%(w/v) agarose gel and purified using Wizard® SV Gel and PCR Clean-up system (Promega, Hampshire, UK) and DNA was eluted in 50µl nuclease-free water.

Restriction Endonuclease Digest

The appropriate vector, either pET24a(+) or pET24a(+)*Nde*I6his, and the purified PCR product, were usually digested with *Nde* I and *Xho* I (NEB, Hitchin, UK) following the recommended protocol. Digested vectors were electrophoresed on an agarose gel alongside uncut vector to assess whether complete digestion had taken place. Once fully digested, the vectors were treated with Antarctic Phosphatase (NEB, Hitchin, UK) to remove the 5' phosphate group, so preventing self-ligation of the plasmid. Antarctic phosphatase reactions were carried out according to NEB instructions. Digest products were purified using Wizard® SV Gel and PCR clean-up system, and run on an agarose gel with 0.5µg of 2-log ladder to allow estimation of DNA concentration.

Ligations

Ligation reactions typically contained 1x T4 DNA ligase reaction buffer, 12.5U T4 DNA Ligase (NEB, Hitchin, UK), 1:4 molar ratio of vector to insert and were made up to 10µl with nuclease-free water. Reactions were incubated overnight at 16ºC and then heated to 70ºC for 20min to denature the ligase enzyme. The DNA was then precipitated using ethanol as detailed in chapter 2 and re-suspended in 5µl of nuclease Milli-Q water to allow electroporation.

Electroporation

E. coli strain KRX (Promega, Southampton, UK), containing the pRARE2 plasmid (Rosetta2, Novagen) that encodes seven rare codons, was made

competent for the uptake of plasmid DNA following the method as detailed in chapter 2. 0.5µl of the ligation product was incubated on ice with 40µl KRX (pRARE2) in 20%(v/v) glycerol for 2min and electroporation carried out as detailed in chapter 2. After the 1h incubation step at 37ºC, 100µl of the cells were spread onto LB agar plates containing kanamycin (50µg/ml) chloramphenicol (34µg/ml) and incubated overnight at 37ºC.

Colonies from the overnight incubated plates were gridded and screened using PCR as detailed in chapter 2 methods. Screening primers used for pET vectors were T7 promoter and T7 terminator sequences (Appendix I).

Mini Plasmid Prep

5ml of LB containing kanamycin and chloramphenicol were inoculated with positive colonies identified from the PCR screen and incubated overnight at 37ºC with aeration at 225rpm. Cultures were then centrifuged at 5,000xg for 10min and the supernatant discarded. Plasmid DNA was extracted from the cell pellet using Wizard® plus SV mini-prep kit (Promega, Hampshire, UK). DNA was eluted in 50µl nuclease-free water and a 1µl aliquot loaded on a 1%(w/v) agarose gel to allow an estimation of the concentration of the purified plasmid. The purified plasmid could be used for sequencing as detailed in chapter 2 methods.

Small Scale Protein Expression

KRX (Promega, Hampshire, UK)

Clones containing a pET24a(+) vector construct with the gene of interest were used to inoculate 3ml LB containing 50µg/ml kanamycin and 34µg/ml chloramphenicol and incubated overnight at 37ºC with aeration at 225rpm. 100µl of the overnight culture was used to inoculate 10ml LB containing

antibiotic as before. Cultures were incubated at 37ºC with aeration until an OD_{600} of 0.4-0.5 was reached. The temperature was then lowered to 15-25ºC before protein expression was induced with 0.1%(w/v) rhamnose (Promega, Hampshire, UK) and 1mM IPTG (Melford, Ipswitch, UK) once the OD_{600} had reached 0.5-0.6. Cultures were then incubated overnight at 15ºC with 225rpm aeration. After overnight incubation, the OD_{600} was measured and the cells harvested by centrifugation at 5,000xg for 10min.

ArcticExpress™ (DE3) RIL (Stratagene, Agilent Technologies UK Ltd, Cheshire, UK)

Colonies positive for the gene of interest within a pET24a(+) vector were grown overnight with 20µg/ml gentamycin (Sigma, Gillingham, UK), 75µg/ml streptomycin (Sigma, Gillingham, UK) and 50µg/ml kanamycin. 250ml LB with appropriate antibiotic was inoculated with 2.5ml of the overnight culture and incubated at 30ºC for 3h at 225rpm. The temperature was then decreased to 12ºC for 10min and then IPTG added to 1mM final concentration. Induced cultures were incubated for a further 24h at 12ºC with aeration at 225rpm. The cells were then harvested at 5,000xg for 5min and the supernatant discarded.

12.5% SDS PAGE

12.5% gels were cast with a running gel containing 3ml running gel buffer [1.5M Tris-base, 0.4%(w/v) SDS], 5ml 30%(v/v) acrylamide (BioRad, Hemel Hampstead, UK) and 4ml Milli-Q water. The gel was set with the addition of 12.5µl TEMED (Sigma, Gillingham, UK) and 50µl 10%(w/v) APS (Melford, Ipswitch, UK). The running gel was topped with a stacking gel containing 2.4ml stacking gel buffer [0.478M Tris-Base, 0.4%(w/v) SDS], 0.9ml 30%(v/v) acrylamide, 3.6ml Milli-Q water and set with the addition of 10µl TEMED and 50µl 10%(w/v) APS.

Protein Sample Preparation

The cell pellets from a 10ml culture were re-suspended in cell lysis buffer D and sonicated for 30s at 50W. The cell debris and insoluble protein material were separated from the soluble proteins by centrifuging at 14,000xg for 5min. The supernatant could be used directly as the soluble protein fraction. The cell pellet was washed with cell lysis buffer D, pelleted as before and then re-suspended in the same volume of cell lysis buffer D. This fraction contained the insoluble proteins.

Protein samples were mixed with an equal volume of protein loading buffer and denatured by heating to 100ºC for 3min. Samples were loaded onto a cast 12.5% SDS PAGE gel and were electrophoresed at 70V in the protein running buffer until past the stacking gel. Once through the stacking gel, the voltage was increased to 240V until the tracker dye reached the base of the gel.

At the end of the run, SDS PAGE gels were rinsed in water and then placed in stain for 30min to allow staining of the protein; destain was then used to remove background dye from the gel.

Site-Directed Mutagenesis (SDM)

This procedure was used to alter individual nucleotides in a gene sequence. SDM primers were designed to contain the desired mutation on both strands situated in the middle of the primers and were between 25-45 bases in length. The GC content of the primers had to be a minimum of 40% and ideally terminated in one or more G or C bases.

SDM primers were used with Phusion® DNA polymerase with the template of circular vector containing the desired gene to be altered. The final reaction components consisted of 1x HF buffer, 0.2mM dNTPs, 10pmol of forward and reverse primers, 25ng plasmid DNA, 1U Phusion® DNA polymerase and made to a final volume of 50µl with nuclease-free water. The reactions were cycled as follows:

98ºC	30s		
98ºC	30s	}	18 cycles
55ºC	1min		
68ºC	6min		
68ºC	7min		

1µl of the reaction was electrophoresed on a 1%(w/v) agarose gel to ensure a successful amplification before the addition of 10U *Dpn* I (NEB, Hitchin, UK) and incubation at 37ºC for 1h to remove the original, non mutated template.

The amplified vector was heat-treated at 80ºC for 20min to denature the *Dpn* I and then purified using Wizard® SV Gel and PCR Clean-up system. The DNA was eluted in 30µl nuclease-free water and 1µl transformed into electrocompetent *E. coli* TOP10F'. 100µl of the transformation was spread onto LB agar containing 50µg/ml kanamycin and incubated overnight at 37ºC.

Positives colonies were isolated using the screening PCR protocol as detailed in chapter 2. Where an RE site had been added or removed using SDM, this PCR could be used in an RE digest with the appropriate enzyme to ensure mutagenesis had taken place. Successful mutagenesis was confirmed by using sequence analysis, which also ensured that no additional mutations had occurred.

Overlap Extension PCR

This technique results in the fusion of two genes into one open reading frame, as detailed in the introduction to this chapter, and allows the translation of a single protein containing both gene products. Four primers had to be designed as follows:

Primer 1: Contained the natural start codon of the 5' gene, followed by another 18-25bp of downstream gene-specific bases. In addition, this primer also contained RE sites upstream of the start codon to allow directional cloning of the final product.

Primer 2: An antisense version of primer 3

Primer 3: Contained 20-25 bases of the end of the 5' gene. The stop codon of the 5' gene was replaced with the start codon of the 3' gene, followed by additional 20-25 downstream gene-specific bases of the 3' gene

Primer 4: Contained 18-25 gene-specific bases ending with the stop codon, altered to TAA where necessary, followed by an RE site allowing directional cloning of the gene.

Two initial PCRs were carried out. The first contained 20ng of template for the 5' Gene, 1x HF buffer, 0.2mM dNTPs, 25pmol primer 1 and primer 2, 1U Phusion® DNA polymerase and the final volume made up to 50µl with nuclease-free water. The second PCR was set up in the same way but using template for the 3' Gene and primers 3 and 4.

The PCRs were cycled as follows:

98℃	30s	
98℃	10s	20 cycles
55℃	10s	
72℃	30s/kb	
72℃	7min	

Samples from the reactions were electrophoresed on an agarose gel to ensure that the correct size product has been produced, and then 10U *Dpn* I (NEB, Hitchin, UK) was added followed by incubation at 37℃ for 2h, ensuring removal of parental DNA template. The reactions were then purified using the Wizard® SV Gel and PCR Clean-up system; the DNA was eluted in 50µl nuclease-free water and could be used as a template in a PCR to fuse the two genes together.

The gene fusion PCR used 1x HF buffer, 50pmol primers 1 and 4, 100ng of each purified product from the first two PCRs, 0.2mM dNTPs, 2U Phusion® DNA polymerase and the volume made up to 100µl with nuclease-free water. The reactions were cycled as before and the product visualized on an agarose gel to ensure fusion had taken place. Gel purification was performed as per manufacturing instructions to remove any contamination from the initial parental template. DNA was eluted in 50µl nuclease-free water.

The purified fused PCR product could then be digested and ligated into an appropriate vector as detailed in the methods in Chapter 2.

3.3 – RESULTS

B. caldovelox IEP Gene Cloning and IEP Expression

Gene-Walking of the IEP gene within *B. caldovelox* revealed 642bp of DNA sequence identical to the recombinase of Bh.Int-like element from *G. kaustophilus*, locus tag GK1355. Rather than continuing to Gene-Walk this strain, it was decided to design cloning primers based on the sequence in the database for this gene. Selected cloning vectors for the expression of the IEP were pET24a(+) and a modified version of this vector, pET24a(+)NdeI6his. This modified vector introduced 6 histidine residues onto the N-terminus of the protein when cloning using the *Nde* I RE recognition site at the start of the gene. However, a restriction map of GK1355-IEP gene revealed an internal *Nde* I site toward the 3' end of the gene (Figure 3.3). In order to use the chosen vectors, it was therefore necessary to clone the IEP in two fragments (Figure 3.3), an upper fragment of 495bp and a lower fragment of 816bp, and primer pairs (Appendix I) were designed accordingly:

Upper Fragment:

- *Forward primer:* B.cv_IEP_F1_start_NdeI primer designed with an initial 6 random nucleotides followed by an *Nde* I site incorporated into the start codon followed by an additional 24 bases of gene-specific sequence.
- *Reverse primer:* B.cv_IEP_R1_EcoRI primer designed to flank and incorporate a naturally-occurring *Eco*R I site found upstream of the *Nde* I site.

Lower Fragment

- *Forward primer:* B.cv_IEP_F2_EcoRI primer was an antisense version of the reverse primer B.cv_IEP_R1_EcoRI, designed to flank and incorporate a naturally-occurring *Eco*R I site

- *Reverse primer:* B.cv_IEP_R2_Stop_SalI contained 23 bases of gene-specific sequence ending with the stop codon, altered to TAA, followed by a *Sal* I site and then additional random nucleotides to ensure RE digestion.

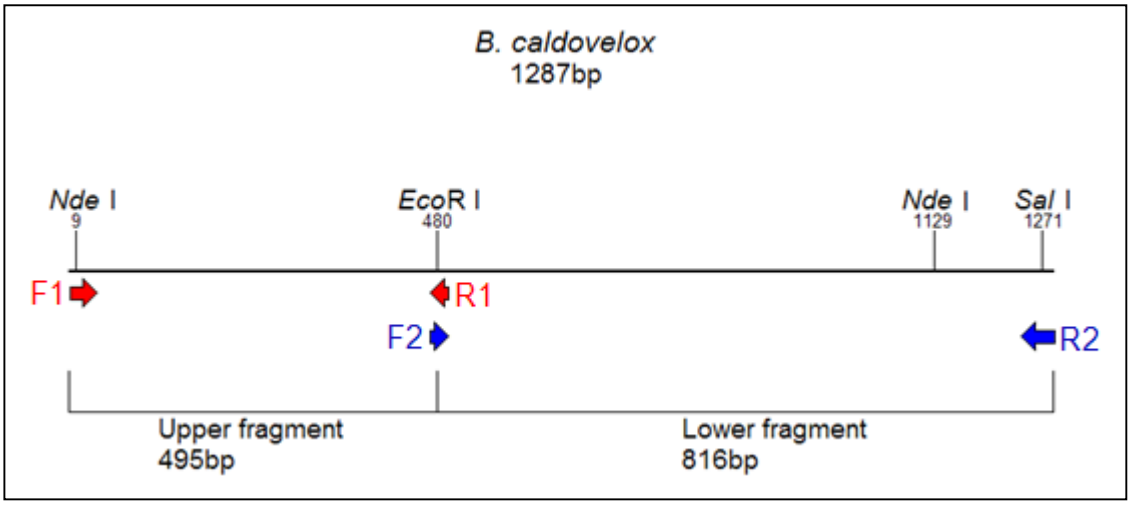

Figure 3.3: A diagram showing the positions of the primers to clone the IEP from *B. caldovelox* into two fragments. The upper fragment can be amplified using the F1 and R1 primer set, and the lower fragment can be amplified with the F2 and R2 primer pair.

Initially the primers were first tested to ensure that they would amplify the expected size fragments and were used with both *B. caldovelox* and *B. caldotenax* gDNA. Interestingly, despite the fact that this IEP gene was only Gene-Walked from *B. caldovelox*, amplification of the correct size product occurred using both strains as template (Figure 3.4). This could be explained by the results of the Touch-Down diagnostic PCR (Figure 3.5).

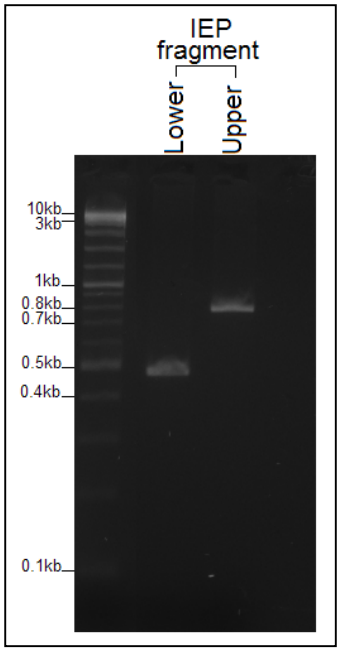

Figure 3.4: A 1%(w/v) agarose gel showing the two IEP fragments amplified from *B. caldotenax* gDNA. The correct sizes of 816bp for the lower fragment and 495bp for the upper fragment were observed.

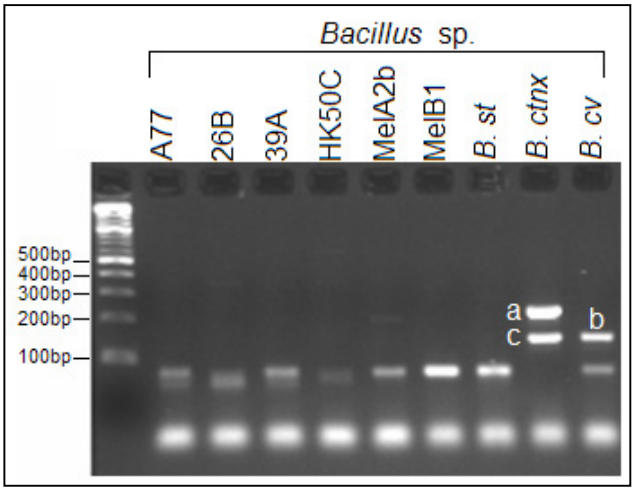

Figure 3.5: Touch-Down PCR results on various strains. *B. caldotenax* contained two possible IEP genes (labelled as a and c). Amplicon (a) from *B. caldotenax* was identified as being from an IEP but subsequent Gene-Walking was unsuccessful. *B. caldovelox* contains one fragment (b) that has been identified as identical to the GK1355-IEP and this amplicon runs at approximately the same size as the unidentified fragment (c) from *B. caldotenax*.

When the initial Touch-Down PCRs were carried out on *B. caldotenax* (Chapter 2) there were in fact two bands between the critical size of 130-300bp. The larger fragment was identified as an RT domain with a different sequence from the now almost fully identified IEP from *B.* caldovelox. However, Gene-Walking failed to obtain the remainder of this gene. The second, smaller fragment with *B. caldotenax* remained unidentified. However, the fact that this smaller fragment runs at approximately the same size as the fragment from the *B. caldovelox* IEP, in addition to the amplification product of correct size being seen with the *B. caldovelox* primers and using *B. caldotenax* gDNA as template (Figure 3.4), suggested that this fragment could correspond to a similar IEP. Therefore, both IEP genes were cloned and sequenced.

The upper and lower fragments from both *Bacillus* species were amplified for cloning using Phusion® DNA polymerase and 50ng of gDNA template. *B. caldovelox* also produced fragments of approximately the correct size by comparison to a marker on an agarose gel (Figure 3.6). The upper fragment was expected to be 495bp and the lower fragment was expected at 816bp.

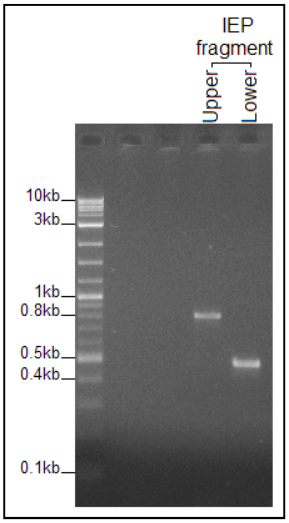

Figure 3.6: An agarose gel showing the sizes of the upper and lower fragments of the IEP genes from *B. caldovelox*.

The PCR products were purified and the upper fragment digested with *Nde* I and *Eco*R I, while the lower fragment was digested with *Eco*R I and *Sal* I. The two vectors, pET24a(+) and pET24a(+)NdeI6his, were digested with *Nde* I and *Sal* I RE and used in a ligation reaction with the two digested fragments. 0.5µl of the heat-treated and ethanol-precipitated ligation was transformed into electrocompetent *E. coli* KRX (pRARE2). 20 colonies from each transformation were then screened by PCR using the T7 promoter and terminator primers to identify those containing the insert of correct size. Screening of the colonies revealed four *B. caldovelox* IEP clones in pET24a(+), three *B. caldovelox* IEP clones in pET24a(+)NdeI6his, and one *B. caldotenax* IEP clone in pET24a(+)NdeI6his.

All positive colonies, along with one negative colony containing just vector with no insert, were selected for protein expression studies. 10ml of LB with appropriate antibiotics was inoculated with 100µl of an overnight culture and incubated at 37ºC with aeration at 225rpm. Protein expression was induced with the addition of rhamnose and IPTG once an OD_{600} of 0.6 was reached. The cultures were incubated overnight (approximately 20h) at 25ºC, 225rpm, and then the cells harvested by centrifugation. The cell pellet was re-suspended in cell lysis buffer D and sonicated. The sample was centrifuged and the supernatant retained as the soluble fraction. The cell debris pellet was washed and re-suspended in 750µl cell lysis buffer D to be used as the insoluble fraction. 5µl of these cell fractions, equivalent to 67µl of the original culture, were electrophoresed on a 12.5%(w/v) SDS PAGE gel (Figure 3.7).

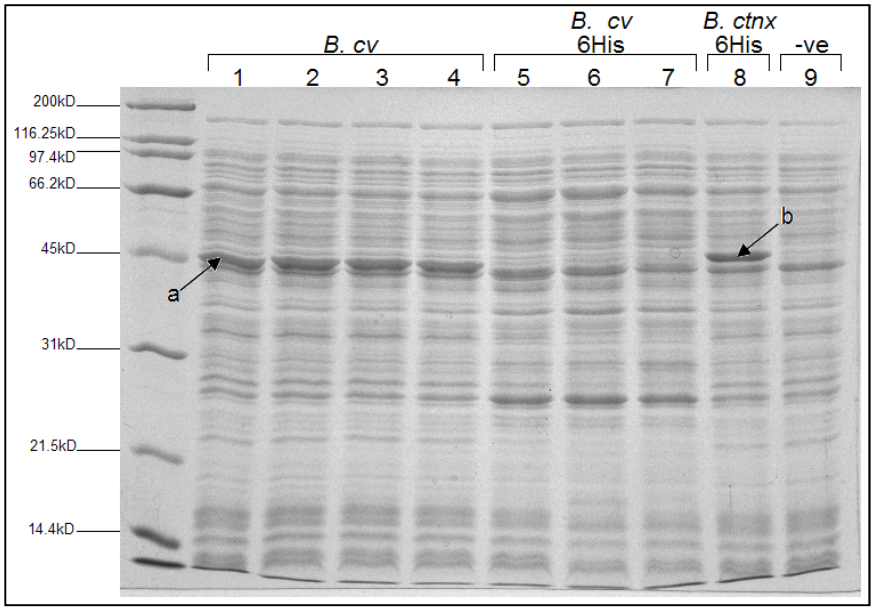

Figure 3.7: 12.5% SDS PAGE gel. Lanes 1-4 show the soluble fraction of the protein expression of the IEP from *B. caldotenax* in pET24a(+); the arrow (a) marks the over-expressed protein expected to be the IEP due to its absence in the negative-control in lane 9. Lanes 5-7 show the protein expression from four clones containing an IEP gene from *B. caldotenax* in pET24a(+)NdeI6his. Lane 8 shows the protein expression from the cloned IEP gene from *B. caldotenax* in pET24a(+)NdeI6his, arrow (b) marks the additional over-expressed protein, expected to be the tagged IEP, not seen in the negative-control. Lane 9 is the protein expression from a pET24a(+) clone negative for any insert.

The protein expression gel showed that clones 1-4 were expressing a protein in addition to those seen in the negative-control lane. This band was running level with the 45kD marker, slightly lower than the expected size of 49kD for this IEP. No over-expressed protein was present in either the soluble or insoluble fraction for *B. caldovelox* IEP genes in pET24a(+)6his, seen in lanes 5-7. The *B. caldotenax* IEP with the N-terminal 6his-tag did have an additional protein band from those seen in the negative-control ran slightly higher than that non-tagged protein suggesting that his-tagging had been successful. In clones successfully expressing the IEP there was a small percentage of the over-expressed protein present in the insoluble fraction and this was increased with the N-terminal his-tagged protein. However, it was found that by inducing protein expression at

the lower temperature of 15ºC, the IEPs from these *Bacillus* species were 100% soluble.

Sequencing analysis of the IEP genes from both *Bacillus* species showed them to be identical to the fully-sequenced *G. kaustophilus* GK1355-IEP. This proved that the additional IEP fragment, previously unidentified in *B. caldotenax* ,was in fact identical to the IEP genes that found in both *B. caldovelox* and *G. kaustophilus*. Given that *G. kaustophilus* is found to have three seperate IEPs, *B. caldotenax* has been identified to have two IEPs and *B. caldovelox* has one IEP it is unlikely that this occurence of the same IEP within all three bacteria is due to a contamination with the same bacteria being studied. It is more likely that this repeated occurence of the same intron is due to the mobile nature of the introns themselves with each different bacterium acquiring the same retroelement containing an identical IEP.

B. caldovelox IEP Manipulation

Since the IEP from *B. caldovelox* was expressing with 100% solubility it was chosen for further manipulation experiments in an attempt to enhance the properties of the enzyme. This involved the fusing of the DNA-binding protein, Sac7d, onto either the C-terminus or the N-terminus of the IEP.

Initially, site-directed mutagenesis (SDM) had to be carried out to remove the internal *Nde* I site towards the 3' end of the gene to utilize an *Nde* I site that had been incorporated into the start of the gene. Primers were designed (Appendix I) to flank this *Nde* I site and to alter the CAT ATG recognition site to CCT ATG. This mutation was silent to avoid alteration of the amino acid residue that would be incorporated on translation. Phusion® DNA polymerase was used along with the two SDM primers to extend from the primer sites around the pET vector

containing the IEP gene. The reaction was then treated with *Dpn* I which ensured the removal of the parental pET vector template whilst leaving the mutated PCR product intact. After purification, 1µl of this mutated DNA was used in a transformation into *E. coli* strain KRX (pRARE2). Transformed colonies were screened for a successful mutation by first running a screening PCR with T7 promoter and terminator. This PCR product was then used directly in an *Nde* I digest to test for the removal of the recognition site. Any PCR products that were digested into two fragments would contain an *Nde* I site and therefore indicated that SDM had been unsuccessful. Positively mutated colonies were then sequenced using the ABI3100 sequencing protocol to ensure that no additional errors had been introduced into the IEP gene sequence.

To create the Sac7d fusion protein, an overlap extension PCR method had to be used. Primers were designed (Appendix I) to incorporate the Sac7d domain onto either the C-terminus or the N-terminus of the protein (Figure 3.8), joined with a 3 amino acid linker, GTV.

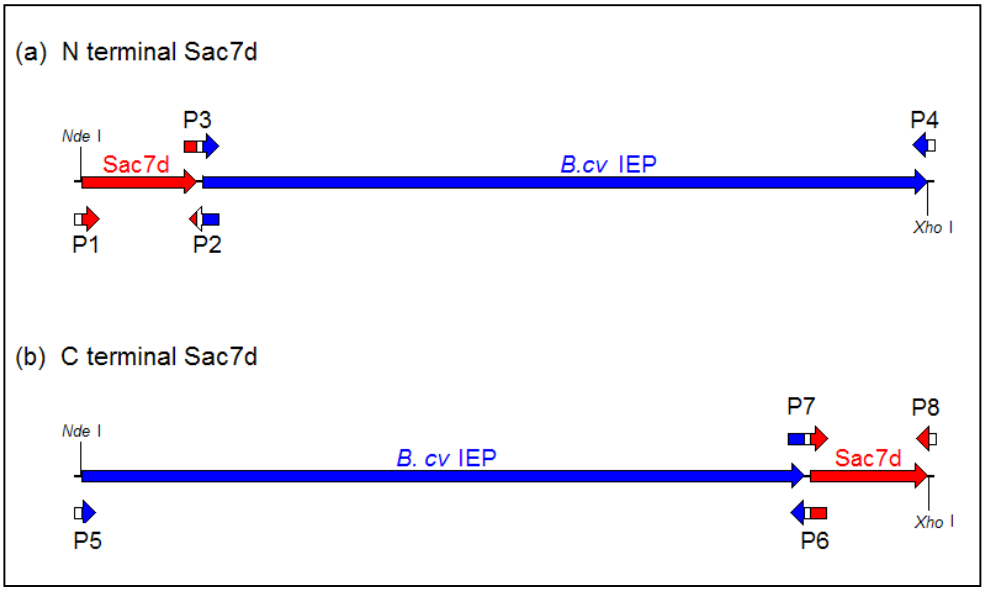

Figure 3.8: A diagram to show the relative positions of the overlap extension PCR to incorporate a Sac7d domain at either the N-terminus or the C-terminus of the *B. caldovelox* IEP. The primers (Appendix I) were as follows:

P1: Sac7d_NterF_*Nde* I
P2: Sac7d_NterR
P3: *B.cv*_CterF
P4: *B.cv*_CterR_*Xho* I
P5: *B.cv*_NterF_*Nde* I
P6: *B.cv*_NterR
P7: Sac7d_CterF
P8: Sac7d_CterR_*Xho* I

The template used to amplify the Sac7d gene was a vector, donated by GeneSys Ltd, containing the Sac7d gene that had been artificially synthesized to allow optimized codon usage for *E. coli* (Appendix III). 2ng of this vector was used in a PCR with Phusion® DNA polymerase with the following primer combinations:

- P1 with P2 where the Sac7d was required at the N-terminus of the fusion protein.
- P7 and P8 where Sac7d was required at the C-terminus of the fusion protein.

The amplified products were visualized on an agarose gel to ensure the correct size of 250bp for the N-terminal Sac7d and 255bp for the C-terminal Sac7d had been achieved (Figure 3.9a).

The cloned *B. caldovelox* IEP gene, with removed internal *Nde* I site, was also amplified with Phusion® DNA polymerase and the following primer combinations:

- P3 and P4 where the *B. caldovelox* IEP was required at the C-terminus of the fusion protein.
- P5 and P6 where the *B. caldovelox* IEP was required at the N-terminus of the fusion protein.

These PCR products were visualized on an agarose gel to ensure that approximately the correct sized products of 1292bp for the C-terminal IEP and 1307bp for the N-terminal IEP were achieved (Figure 3.9b).

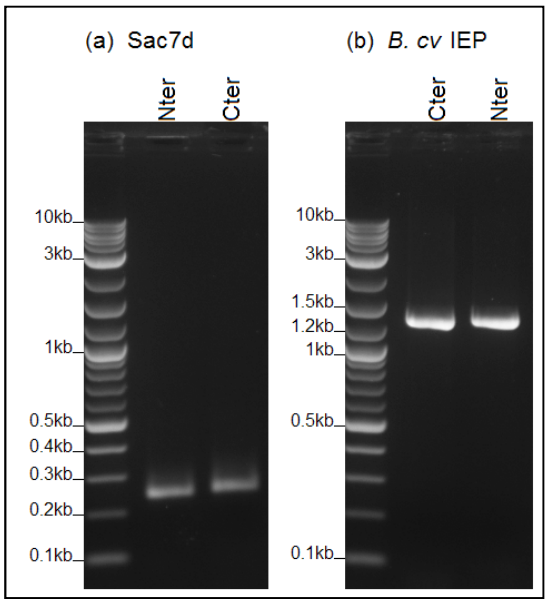

Figure 3.9: An agarose gel showing Phusion® DNA polymerase amplified products of:

(a) Sac7d gene for overlap extension PCR for both N and C-terminal fusion products.

(b) *B. caldovelox* IEP gene product for both C and N-terminal fusion products.

Once confirmed as being approximately the correct size, the four PCR products were treated with *Dpn* I to remove parental plasmid template, and then purified. Overlap extension PCRs were set up with 100ng of each template as follows:

- To create Sac7d-IEP fusion product: N-terminal Sac7d PCR product and C-terminal IEP PCR product used as template with primer pair P1 and P4.
- To create IEP-Sac7d fusion product: N-terminal IEP PCR product with C-terminal Sac7d PCR product with primer pair P5 and P8.

The product sizes from these overlap extension PCRs were expected to be 1495bp. Gel purification had to be carried out to ensure only the correct size product was present (Figure 3.10).

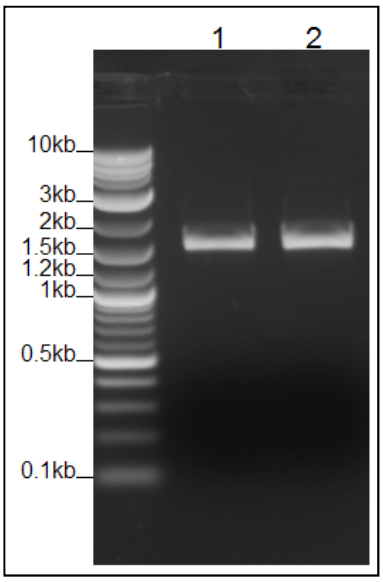

Figure 3.10: A 1% (w/v) agarose gel showing the overlap extension PCR products. Lane 1 contains the IEP gene with a C-terminal Sac7d domain (IEP-Sac7d) and lane 2 shows the IEP gene with an N-terminal Sac7d domain (Sac7d-IEP).

The purified overlap extension PCR products were digested with REs *Nde* I and *Xho* I and ligated into pET24a(+). Once the ligation reactions had been heat-treated and ethanol-precipitated, 0.5µl was transformed into

electrocompetent *E. coli* KRX (pRARE2) and 100µl plated onto LB plates with kanamycin and chloramphenicol. Colonies were screened using a colony-screening PCR and T7 promoter and terminator primers, colonies giving the correct size PCR product were assumed to be positive for the fusion protein.

Positive colonies were selected for protein expression studies along with a pET24a(+) vector containing no insert to be used as a negative-control. Cells were harvested from 10ml LB culture expressed at 15ºC with a 20h rhamnose and IPTG induction. The cell pellet was re-suspended in 750µl cell lysis buffer D and then sonicated. 5µl of both the soluble and insoluble fractions, equivalent to 67µl of original culture, were electrophoreses on a protein gel (Figure 3.11).

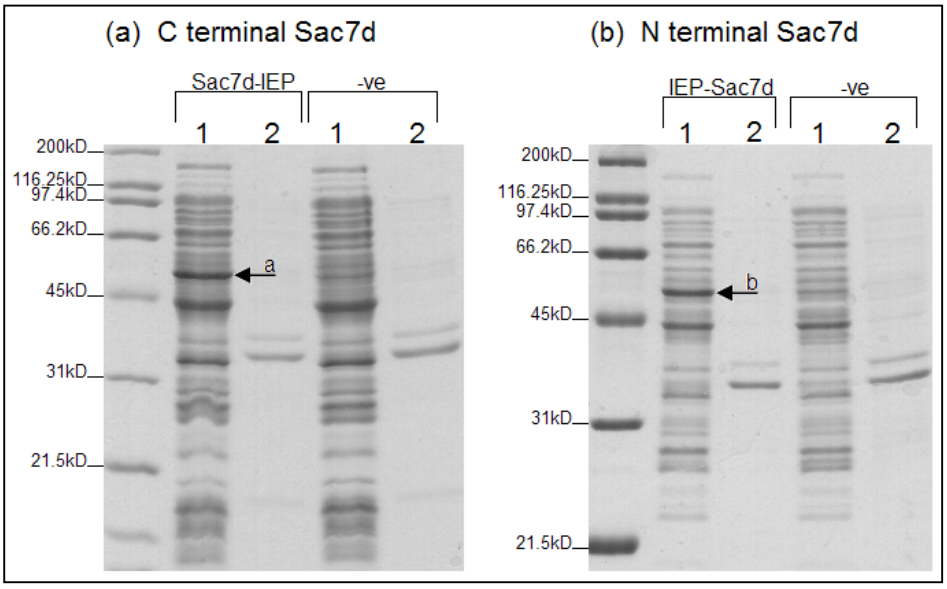

Figure 3.11: 12.5%(v/v) SDS PAGE gels showing the soluble (1) and insoluble fraction (2) of:

(a) The expression of the fusion product with Sac7d at the C-terminus with arrow (a) marking the over-expressed fusion product not seen in the negative-control lane (-ve).

(b) The expression of fusion protein with Sac7d at the N-terminal of the IEP with arrow (b) showing the over-expressed protein not seen in the negative-control lane (-ve).

Both Sac7d fusion clones showed over-expression of a protein that was not present in the negative-control. By plotting log Mr values of the size standard against the distance travelled it was possible to estimate the protein size at 52kD. This was slightly lower than the expected 57kD as predicted from the protein sequence. No over-expressed protein was present in the insoluble fraction suggesting that the protein was 100% soluble. DNA sequence analysis of these fused genes showed a successful read through in one ORF and no errors present in either the Sac7d or the IEP domain.

Thermosinus carboxydivorans Gene Cloning and Protein Expression

A BLAST search for IEPs revealed *T. carboxydivorans* to contain a fully sequenced IEP, from a draft genome, with 63% protein sequence identity to that of *G. kaustophilus* HTA426 GK1355-IEP. Cloning primers (Appendix I) were designed to amplify this gene to allow its cloning into a vector and its subsequent protein expression. Two lower primers were designed, one to alter the stop codon from TGA to TAA making it a stronger stop codon in *E. coli* and then finishing in an *Xho* I recognition site, and the second primer to replace the stop codon with an *Xho* I site so that cloning into pET24a(+) allowed the addition of a C-terminal his-tag.

T. carboxydivorans, DSMZ Number 14886, was ordered from the DSMZ as a lyophilized sample. The gDNA was extracted using the method by Götz *et al*. (2002), detailed in Chapter 2, and then the yield further enhanced by using GenomiPhi™V2 DNA amplification kit. A Phusion® PCR was set up to amplify the gene using both primer sets. The PCR product was electrophoresed on a 1%(w/v) agarose gel to ensure the correct-sized product had been amplified (Figure 3.12). A size of 1248bp was expected for the C-terminal his-tagging product and 1251bp for the product not to be his-tagged or for N-terminal his-tagging.

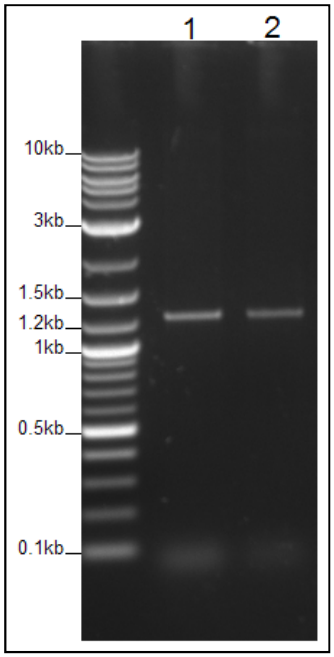

Figure 3.12: An agarose gel of the IEP gene amplified from *T. carboxydivorans*. Lane 1 shows the product that will allow N-terminal his-tagging and non-his tagging of the protein and lane 2 shows the product that will enable C-terminal his-tagging.

Once the correct size was confirmed, the PCR product was digested with *Nde* I and *Xho* I and ligated into either pET24a(+) or pET24a(+)NdeI6his that had been digested with the same REs. Heat-treated and ethanol-precipitated ligation products were transformed into electrocompetent *E. coli* KRX (pRARE2) and plated onto LB agar plates with kanamycin and chloramphenicol. Colonies were screened by PCR and assumed as positive by the presence of a band of the correct size on an agarose gel.

Positive colonies were selected for gene expression studies along with one negative colony. Protein expression was induced according to KRX protocol instructions with induction temperatures set at 15ºC. The cell pellet from a 10ml induced culture was re-suspended in 750µl of cell lysis buffer D and sonicated. 5µl, equivalent to 67µl of original culture, was electrophoresed on a protein gel

for expression analysis. Of the three versions of the IEP gene cloned, the non-tagged, the C-terminal his-tag and the N-terminal his-tag, only the C-terminal his-tagged IEP showed any over-expression of a protein different from that in the negative-control (Figure 3.13).

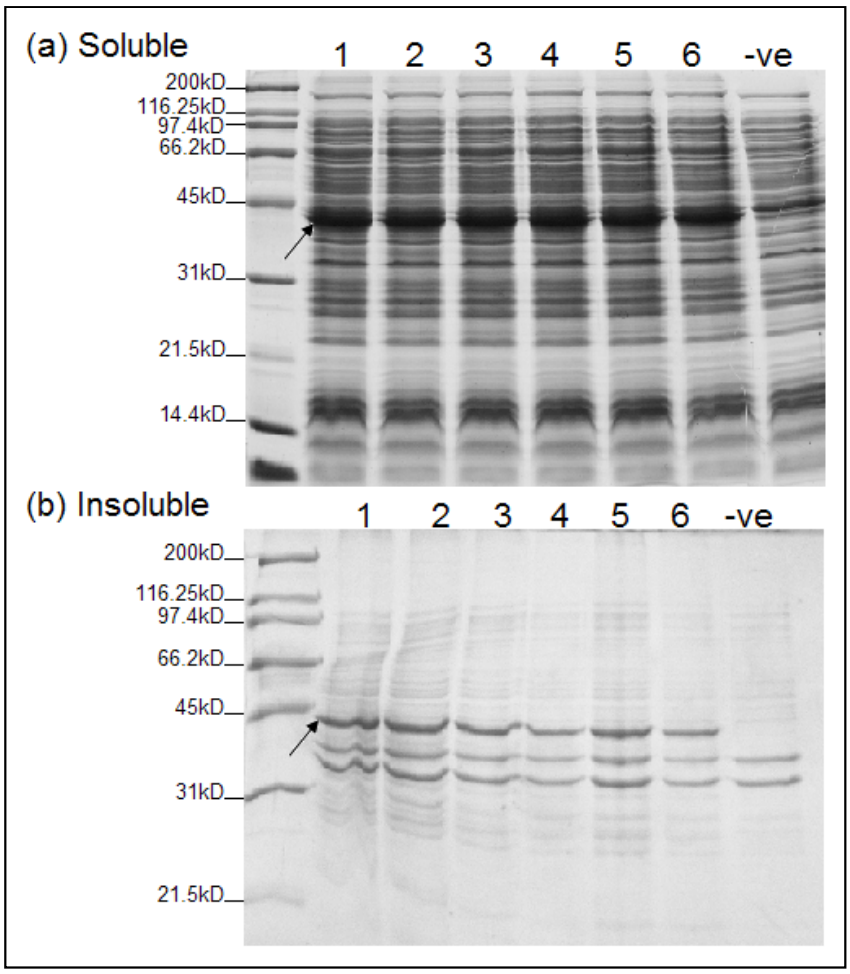

Figure 3.13: 12.5% SDS PAGE gels showing:
(a) The soluble fraction of the over-expression of *T. carboxydivorans* IEP.
(b) The insoluble fraction. The arrow marks the expected IEP protein band present in all clones (lane 1-6) and which is not present in the negative-control lane (-ve).

The IEP gene in pET24a(+)NdeI6his vector clone had an additional over-expressed protein, not present in the negative-control lane, which was

estimated by using the marker standards to be 43kD, slightly lower than the expected size of 47Kd predicted from the protein sequence. This extra band, suggestive of the IEP was also present, at low levels, in the insoluble fraction (Figure 3.13b).

To achieve a maximum yield of soluble protein, an attempt was made to reduce the level of the insoluble fraction. Initially, rhamnose concentrations were reduced; however, this resulted in a reduction in the level of protein expressed with no affect on the ratio of soluble to insoluble protein. Another option was to try a different *E. coli* expression strain and ArcticExpress™ (DE3) RIL was selected as an alternative. Positive colonies were grown overnight in 5ml LB with appropriate antibiotic and the plasmid purified using the Promega Wizard® Plus SV mini-prep DNA purification system. Purified plasmids were transformed into electrocompetent ArcticExpress™ (DE3) RIL and plated on LB with 50µg/ml kanamycin, 20µg/ml gentamycin and 75µg/ml streptomycin. ArcticExpress™ colonies were then used to inoculate 250ml LB with antibiotic and protein expression induced according to the protocol. A 10ml sample from the 250ml culture was collected and the cells harvested to analyse for protein expression. Protein expression using ArcticExpress™ (DE3) RIL cells was lower than that seen from KRX (pRARE2) with no change in the ratio of insoluble protein. Therefore, with higher a higher expression yield, KRX (pRARE2) would be used with the standard expression protocol when this IEP was to be expressed on a larger scale.

Petrotoga mobilis IEP Gene Cloning and Protein Expression

The three IEP genes identified from *P. mobilis* were very similar in sequence to each other with IEP1 and 2 being 98% identical, 2 and 3 being 95% identical and 1 and 3 being 94% identical (Appendix IV). However, these slight differences on a DNA level did alter the amino acid composition (appendix IV)

so attempts were made to clone all three IEPs. Primers were designed (Appendix I) to try to amplify the individual genes by making use of DNA base changes at the last nucleotide of the 3' end of the primers. This allowed the design of two forward primers, one for IEP1 and 2 and a separate one for IEP3 and three individual reverse primers, each with a different 3' end nucleotide. It was hoped that these primers would be selective enough to allow the amplification of all three individual IEPs.

P. mobilis, DMSZ number 10674, was ordered from the DSMZ as a lyophilized sample. The gDNA was extracted and the yield enhanced using GenomiPhi™V2 DNA amplification kit. Phusion® DNA polymerase was used to amplify the three IEPs in separate reactions. The PCR products were then gel purified to remove non-specific products and analysed on an agarose gel to ensure the correct size of 1452bp, for each IEP, had been achieved (Figure 3.14).

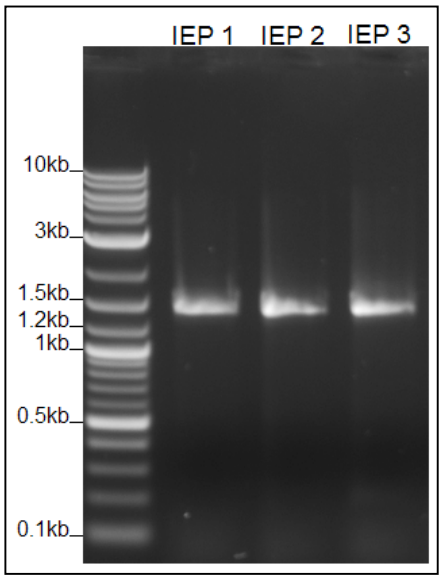

Figure 3.14: A 1%(w/v) agarose gel showing the PCR products from the *P. mobilis* IEP gene amplification. The products are running at approximately the expected size of 1452bp.

The IEP gene PCR products were digested with *Nde* I and *Xho* I and ligated into pET24a(+) vector. 0.5µl of ligated, ethanol-precipitated reactions were transformed into electrocompetent KRX (pRARE2) and plated onto LB plates with antibiotic. Colonies were screened using colony PCR and positive colonies containing the correct-size insert were selected for protein expression studies. The cell pellet from 10ml LB cultures, with protein expression induced at 15ºC with IPTG and rhamnose, were harvested and re-suspended in 750µl cell lysis buffer D. 5µl of both the soluble (Figure 3.15, 3.16 and 3.17) and insoluble protein fractions were loaded on to a protein gel

.

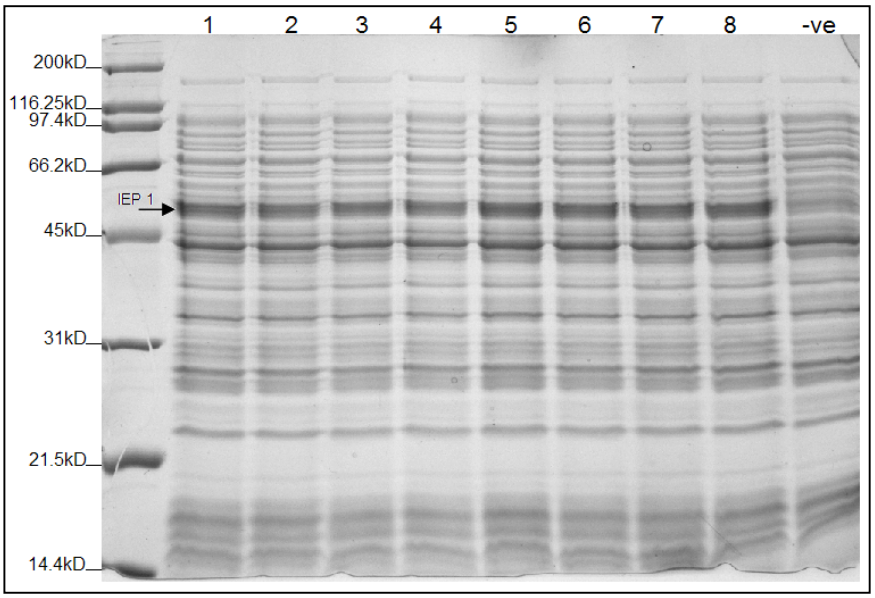

Figure 3.15: 12.5%(v/v) SDS PAGE gel showing the soluble fraction of the protein induction of *P. mobilis* IEP 1 in pET24a(+). All 8 clones are positive for an over-expressed protein not seen in the negative-control lane (-ve). The arrow on the gel marks the over-expressed protein.

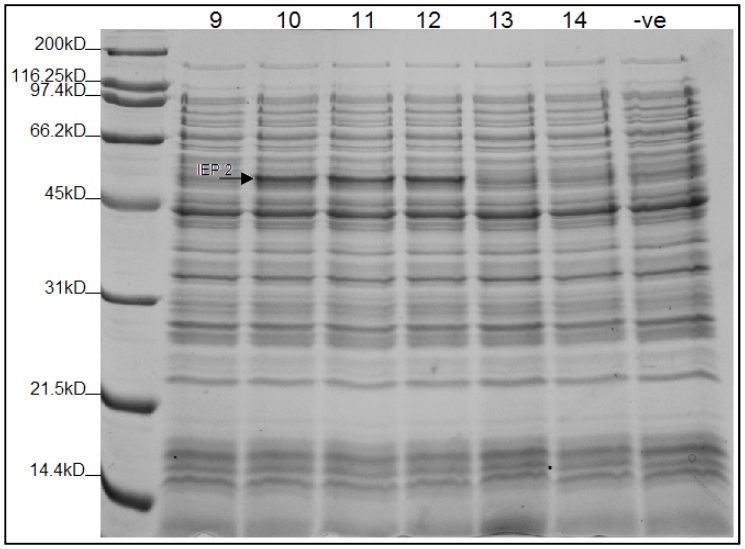

Figure 3.16: 12.5%(v/v) SDS PAGE gel showing the soluble fraction of the protein induction of *P. mobilis* IEP 2 in pET24a(+). Three of the colonies, numbers 10, 11 and 12 showed positive for an extra over-expressed protein not present in the negative-control lane (-ve). The arrow on the gel marks the over-expressed protein.

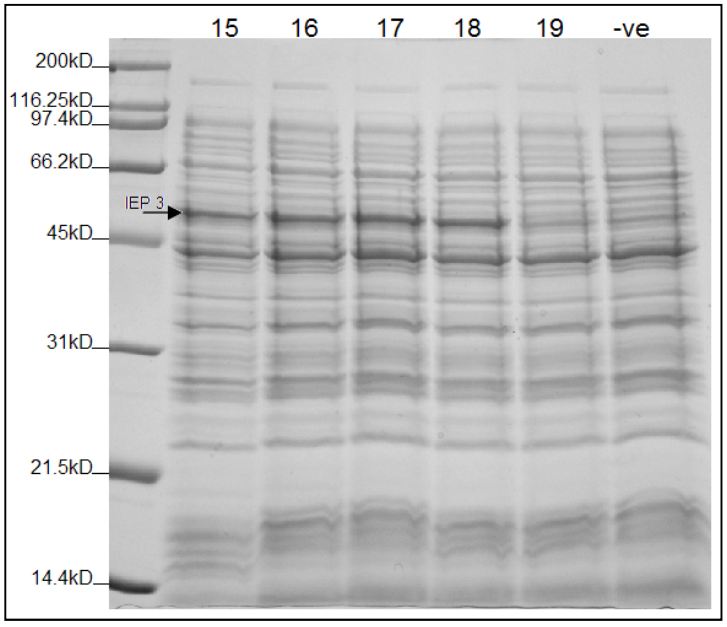

Figure 3.17: 12.5% SDS(v/v) PAGE gel showing the soluble fraction of the protein induction of *P. mobilis* IEP in pET24a(+). Four of the colonies, numbers 15, 16, 17 and 18 showed positive for an extra over-expressed protein not present in the negative-control lane (-ve). The arrow on the gel marks the over-expressed protein.

The expression studies of all three *P. mobilis* IEPs showed a positive band not present in the negative-control lane. Plotting the markers as standard size against distance travelled, this additional protein band was estimated to be 52kD in size, very slightly lower than the expected protein size of 54kD predicted from the IEP amino acid sequence. The positive colonies were grown on a 5ml scale and the plasmid purified using the Promega mini-prep kit. The eluted plasmid was sequenced using the ABI 3100 sequencing protocol to verify the correct sequence was present. All the clones sequenced for IEP1 and 2 were correct. However, the sequencing results showed that out all the possible IEP3 clones, only the clone not expressing any soluble or insoluble protein, clone 19, was in fact IEP3. Sequencing analysis of the other clones from this expression trial revealed them all to be identical to IEP2. This was most likely explained by the nature of the primers used in the gene amplification. The fact that the specific primers for each IEP differed by very few bases led to the possibility of them annealing to the incorrect gene resulting in the wrong gene being cloned and therefore the wrong protein expression. The lack of actual expression of the correct IEP3 could not be explained as sequencing revealed it to be correct. No further work was carried out to express the IEP3.

New Zealand Strain 11 *B. stearothermophilus* Gene Cloning and Protein Expression

The full Gene-Walking of the IEP from the New Zealand strain 11 *B. stearothermophilus* was not successfully completed. However, an alignment using ClustalX showed 96% DNA sequence similarity to the *B. stearothermophilus trt* gene. Cloning primers were therefore based on this fully-sequenced gene and designed in such a way to ensure that the terminal 3' base of the primer did not end in the third base of an amino acid codon as this was the most likely base to differ across different strains. Phusion® DNA polymerase was used to amplify the IEP from strain 11 and the product analysed on an

agarose gel (Figure 3.18). Based on the *trt* gene, the product was expected to be 1294bp in length.

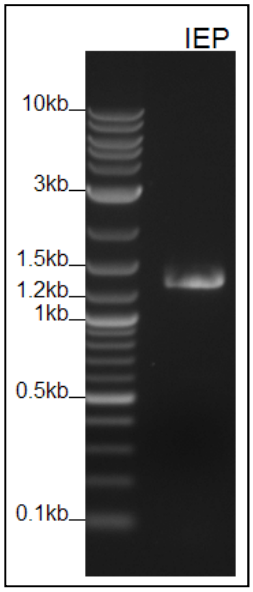

Figure 3.18: A 1%(w/v) agarose gel showing the amplified IEP gene product from a New Zealand strain of *B. stearothermophilus*.

The amplified IEP gene product was purified and digested with *Nde* I and *Xho* I and ligated into pET24a(+). The overnight ligation reaction was then heat-treated, to denature the ligase, ethanol-precipitated and 0.5µl transformed into electrocompetent KRX (pRARE2). Transformed colonies were screened using colony PCR and positives selected for protein expression studies.

The cell pellets from positive colonies and one negative-control were harvested from 10ml cultures after overnight expression of the protein with IPTG and rhamnose at 15ºC. These pellets were re-suspended in 750µl cell lysis buffer D and 5µl of the soluble and insoluble cell fraction was loaded on to a protein gel (Figure 3.19).

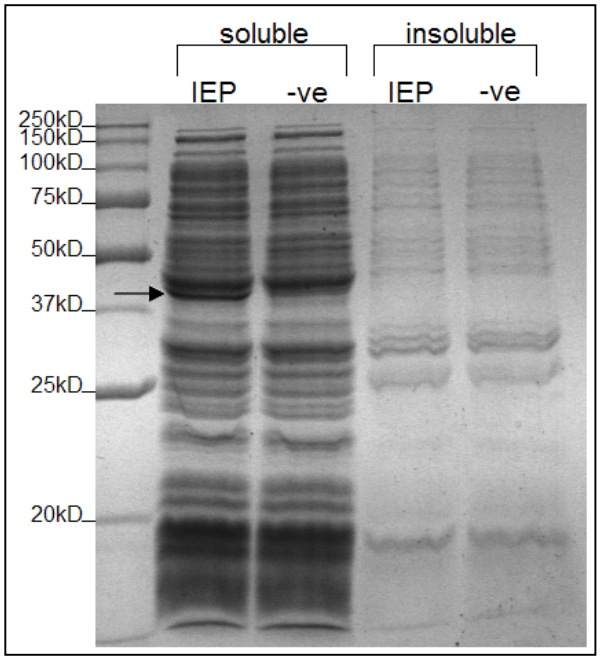

Figure 3.19: A 12.5% SDS PAGE gel showing over-expressed IEP from a New Zealand *B. stearothermophilus* strain. The arrow marks the over-expressed protein not present in the negative-control lane (-ve).

The protein gel showed the presence of a band that was absent from the negative control. Plotting log Mr values of the marker standard against the distance travelled this protein was estimated to be 39kD in size. This was lower than the expected 48kD protein predicted based on the Trt protein. The colony that was over-expressing a protein not present in the negative-control was sequenced using the ABI3100 sequencing protocol to discover the entire IEP gene sequence that was yet to be fully identified. The sequencing analysis revealed that this New Zealand strain of *B. stearothermophilus* IEP gene had 96% sequence identity to the *trt* gene and 97% identify once translated into a protein sequence (Appendix IV). The discrepancy that is seen with this IEP could be due to an incorrect protein being expressed. Although sequencing reveals an IEP within this clone problems are encountered, and discussed later on in this report, with the large scale expression of this protein. Since the protein is the incorrect size and is difficult to express, mass spectrometry analysis would be required to verify if this protein is indeed correct.

3.4 – DISCUSSION

The previous chapter outlined the identification of IEPs within different thermophiles. This chapter outlines their successful cloning and expression even in situations where the full sequence had not been fully identified.

Primers based on the GK1355-IEP from *G. kaustophilus* allowed the amplification of an identical gene within *B. caldovelox*. It was also revealed that the additional IEP within *B. caldotenax*, that was previously unidentified, was capable of being amplified using these same primers and was revealed to be identical to GK1355-IEP. Gene expression studies of these IEPs revealed an over-expressed protein running at approximately 45kD and which was not present in the negative-control lane. This was expected to be the IEP although the protein size was lower than the expected size of 49kD. This protein was 100% soluble when using KRX (pRARE2), providing expression temperatures were maintained at 15ºC.

The successful protein expression of the IEP from *B. caldovelox* allowed further manipulations to the protein. Sac7d, a dsDNA-binding protein, was selected for the possibility of enhancing the enzyme. It was hoped that the addition of a Sac7d domain would enhance the processivity of the enzyme and reduce the amount of truncated products, which can be a problem with mesophilic RTs. Overlap extension PCR was used to create a fusion product of a codon optimized Sac7d at either the N- or C-terminus of the IEP. This fusion experiment was successful, allowing two genes to occur in a single ORF. The proteins were tested for expression using KRX (pRARE2). Protein expression studies of these two fusion products revealed an over-expressed protein, running higher than that of the IEP on its own, at approximately the correct size of 57kD. Having an N-terminal or C-terminal Sac7d domain fused to the IEP

seemed to have no effect on the folding of the protein, with 100% solubility being seen using 15ºC protein induction temperatures.

Chapter 2 showed that a BLAST search, using the *G. kaustophilus* HTA426 GK1355-IEP as the query sequence, revealed two thermophiles that contained fully sequenced IEPs: *T. carboxydivorans* and *P. mobilis*.

T. carboxydivorans, which has an optimum growth temperature of 55ºC, contained an IEP with 63% amino acid identify to the GK1355-IEP. Primers were designed based on the sequence in the database and allowed the successful cloning of this IEP gene. Protein expression studies using KRX (pRARE2) with 15ºC induction temperature, revealed the over-expression of a protein, of approximately 60% solubility, running at a size slightly lower than the expected 47kD. Sequencing of these clones revealed that the IEP contained no mutations. Attempts were made to reduce the level of insoluble protein, including varying the levels of rhamnose and altering the expression strain to ArcticExpress™(DE3) RIL. However, no improvements were seen in the ratio of soluble protein to insoluble. Although the protein was at approximately 60% solubility, the overall expression yield of the protein was in fact high so it was deemed unnecessary to pursue any further work to enhance solubility as enough protein would be present for purification.

P. mobilis, with an optimum growth temperature of 55ºC, was found to contain three IEPs, two with 54% amino acid identity and one with 53% amino acid identity to the GK1355-IEP within *G. kaustophilus*. Cloning primers were designed to amplify all three IEPs allowing their insertion into a pET vector. Colonies positive for an insert were grown up for protein expression studies revealing the presence of a protein, with 100% solubility, at approximately the

correct size of 54kD. However, sequencing analysis revealed that only IEP1 and 2 were successfully expressing and, despite the fact that there were no errors in the sequence, IEP3 failed to be successfully expressed.

A *B. stearothermophilus* strain from New Zealand was revealed to contain an IEP. Gene-Walking failed to completely identify the full-length IEP gene sequence; however, the similarity of the gene to the *B. stearothermophilus trt* gene allowed the design of cloning primers. The IEP gene from this New Zealand strain was successfully amplified, allowing its insertion into a pET vector. Protein expression was carried out using KRX (pRARE2) with a 15ºC expression temperature. The protein expression revealed the presence of an over-expressed protein, not present in the negative-control lane, running a 10kD lower than the expected size at approximalty 39kD. Sequence analysis of this gene revealed it to have 96% identity to the *trt* gene with 97% identify to the Trt protein.

This report has shown the identification of several different IEPs from thermophiles with growth temperatures ranging from 55-70ºC. Two groups have shown that an IEP from *B. caldolyticus* EA1 and the Trt from *B. stearothermophilus* are expressed in low yields with a very percentage of insolubility, leading to difficulty in the purification and subsequent enzyme characterisation (Ng *et al. 2007* and Vellore *et al*. 2004). Unlike these previous studies on IEPs, the IEPs from this report have high expression levels and only the IEP from *T. carboxydivorans* showing some degree of insolubility. These high yields of protein should enable the purification and characterization of the IEPs as RT enzymes.

Chapter 4 – Intron-Encoded Protein Purification

4.1 – INTRODUCTION

Unlike many other enzymes, the RT activity of IEPs could not be assayed directly from cell extracts. This is due to the fact that un-fractionated extracts are likely to contain contaminating RNases and DNases naturally produced by the *E. coli* expression strain. The RNases will degrade the RNA template required for cDNA synthesis, while the DNases will degrade both the cDNA produced and inhibit any subsequent PCR steps. DNases pose less of a problem where heat treatment of the cell extract can be applied. For example, commercially available DNases such as DNase I are denatured by heating at 70ºC for 10min as seen in the manufacturing protocol (NEB, DNase I [online]). RNases, however, can be more problematic to inactivate. RNases are notoriously heat stable; for example, RNase A, while not exhibiting any activity at 90ºC, will show activity once the reaction is quickly cooled to 25ºC (Zale and Klibanov 1986). There are also reports of RNases being able to withstand autoclaving at 121ºC, 15psi for 15min (Pasloske and William, 1998). Purification of the IEP from these contaminants will therefore be necessary before assaying for activity and characterisation of the RT. Several different purification approaches can be adopted to purify the IEP and will be discussed in this introduction

Affinity Chromatography – Ni^{2+} Charged Column

The his-tag incorporated on the N- and C-termini of some of the IEPs should provide a simple purification step. The his-tag, if exposed on the surface of the protein, should bind to a Sepharose column that has been charged with Ni^{2+}

allowing non-tagged proteins to pass straight through. Elution can then be carried out by an imidazole gradient, allowing pure protein to be achieved.

Affinity Chromatography – Heparin Column

This column will make use of the nucleic acid binding properties of the IEP. This protein, along with other nucleic acid binding proteins in the protein sample should bind to the column whilst the other *E. coli* proteins will pass through. Elution can be carried out with an NaCl gradient in an attempt to separate the IEP from the other nucleic acid binding proteins. However, a heparin column can also bind contaminating nucleases, which could potentially be eluted with the IEP. An additional problem could occur if the IEP has a higher affinity to the nucleic acids in the sample over the affinity to the heparin column, a situation that will prevent the IEP from binding. Nucleic acid removal might be required before use of this column as a purification step.

Ion Exchange Chromatography – Cation Exchange Column

The high isoelectric point (pI) of all the IEPs in this report, ranging from 9.8-10.6, pose an advantage for purification from background *E. coli* proteins. Most *E. coli* proteins have a theoretical pI value around 6 (Figure 4.1) and therefore will not be positively charged at pH6.8. At pH6.8 the IEPs should bind the cation exchange column while the majority of the *E. coli* proteins come straight through or elute at low NaCl concentrations.

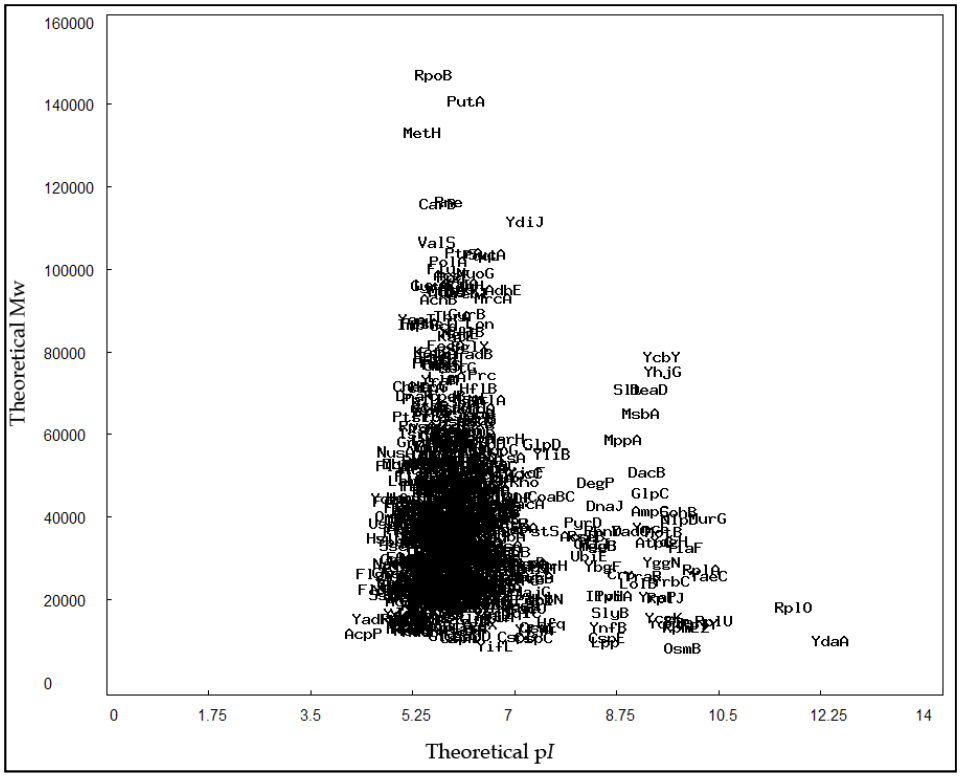

Figure 4.1: A scatter graph plotting the theoretical Mw against the theoretical pI values of *E. coli* proteins. The majority of the proteins have a theoretical pI value of 6. (EcoProDB 2006 [online]).

Ion Exchange Chromatography – Anion Exchange Column

Due to their high pI values it is unlikely the IEPs will bind to this column at pH8. However, the majority of the *E. coli* proteins should bind, along with the negatively-charged nucleic acids. This method could serve as a purification step whereby the nucleic acids are removed from the protein sample along with many of the contaminating proteins, and the IEP should be present in the flow-through.

Ceramic Hydroxyapatite Column (CHT)

Type I CHT, with a higher binding capacity for proteins than Type II, offers an alternative purification method. CHT should offer a method to separate the IEP from nucleic acids, but where this fails a second protocol can be followed. In some situations nucleic acids may have to be removed from the sample, involving high NaCl concentrations, before purification steps can be carried out. Lowering the salt concentration runs the risk of the IEP and other proteins becoming insoluble. Since the CHT is unaffected by high salts, the nucleic acid precipitated sample can be loaded directly onto the column and the NaCl maintained throughout purification, eliminating the risk of precipitation.

Nucleic Acid Removal

Nucleic acids will also need to be absent from the final purified enzyme sample. This will be necessary to avoid any contamination of reactions used to assay the enzyme and to reduce the likelihood of non-specific cDNA synthesis and false results. If the protein purification steps do not separately remove the nucleic acids, or the nucleic acids interfere with the purification, then the removal of nucleic acids may have to be carried out as an initial step prior to downstream purification methods. The use of DNases and RNases would need to be avoided to prevent contamination of the protein sample with these enzymes, which will inhibit subsequent reactions.

4.2 – MATERIALS AND METHODS

Large-Scale Protein Expression

Terrific Broth (TB): 24g yeast extract, 12g tryptone and 4ml glycerol in 900ml Milli-Q water. Once autoclaved 100ml sterile 0.17M KH_2PO_4, 0.72M K_2HPO_4 (Fisher Scientific, Leicestershire, UK) was added prior to use.

4x 2 litre baffled shake flasks containing 1 litre of either LB or TB media were inoculated with 10ml of overnight culture. The cultures were incubated at 37ºC at 225rpm until an OD_{600} of 0.4-0.5 (0.8-1.0 for TB) was reached. The temperature was then reduced to 15ºC and once an OD_{600} of 0.5-0.6 (1.0-1.5 for TB) was reached protein expression was induced with the addition of 0.1%(w/v) rhamnose and 1mM IPTG. Cultures were incubated for approximately 20h at 15ºC with aeration at 225rpm. A 10ml sample was removed from the culture to verify protein expression and the remainder of the cultures were centrifuged at 5,000xg, to pellet the cells, which were then stored at -70ºC until required.

Heat-Treatment of Un-Fractionated Cell Extract

A 10ml sample from the large-scale protein expression was removed and the cells harvested by centrifugation at 5,000xg for 10min. The sample was re-suspended in cell lysis buffer D (chapter 3) and sonicated on ice for 30s at 15micron. After centrifugation at 14,000xg for 5min 100µl aliquots were incubated at 50℃ for 0, 5, 10 and 15min and centrifuged as before. 10µl of the soluble fractions were electrophoresed on a 12.5%(v/v) SDS PAGE gel to verify the presence of soluble IEP before and after heat treatment.

Polyethylenimine Precipitation

Polyethylenimine (PEI) (Sigma, Gillingham, UK) was used to precipitate contaminating nucleic acids from cell extracts where they would otherwise interfere with purification methods. The level of PEI to be used had to be determined for each new cell extract.

In order to test levels required, typically 0.2g of cell pellet was re-suspended in 2ml of a column bind buffer [mentioned below (with 25mM NaCl)]. The sample was sonicated for 30s at 15microns. 100µl aliquots of the sonicated sample were removed and treated with varying concentrations of PEI ranging from 0-0.6%(v/v) and incubated on ice for 1h. The samples were then centrifuged at 14,000xg for 10min. 10µl of the supernatant was electrophoresed on an agarose gel to verify the level of PEI required to precipitate all contaminating nucleic acids. A 12.5%(v/v) SDS PAGE gel was also run in parallel to analyse the level of protein that remains soluble at the required PEI level, as proteins often precipitate out at higher PEI concentrations.

If the protein became insoluble at the required PEI concentrations then additional experiments had to be carried out. The PEI precipitation protocol was repeated using the confirmed level of PEI for that cell extract but with altering NaCl concentrations. Incubation and centrifugation were carried out as before and a 12.5%(v/v) SDS PAGE gel used to confirm the level of NaCl required to allow the protein to stay in solution, whilst an agarose gel ensured all nucleic acids had still been removed.

Once the correct PEI and NaCl levels were determined, the reaction could be scaled up to remove all nucleic acids from a protein sample before loading on to an appropriate column.

Protein Sample Preparation for Purification

With PEI treatment

Cell pellets were re-suspended in the appropriate bind buffer equivalent to 0.1mg/ml of cells. Normally 2g of cells were re-suspended in 20ml of buffer and sonicated for 3x 1min. The total protein fraction was exposed to the appropriate level of PEI, as per the previous protocol, and then centrifuged at 14,000xg for 10min. The supernatant was filtered using a 0.22µm Millipore syringe filter (Fisher Scientific, Leicestershire, UK) and was ready to load on the appropriate column in the same buffer.

Without PEI treatment

The cell pellet from a 250ml expression culture was re-suspended in 5ml of the appropriate bind buffer. The sample was sonicated for 3x 1min on ice and then centrifuged at 14,000xg for 20min. The supernatant was filtered using a 0.22µm Millipore syringe filter and was ready to load on the appropriate column in the same buffer.

If partially purified proteins were to be loaded onto an additional column, the fractions containing the correct protein were pooled and dialysed against the new buffer required for the subsequent purification step.

Bench-Top His Purification

His-bind buffer A: 50mM Tris-HCl pH8.0, 300mM NaCl, 20mM imidazole (Fisher Scientific, Leicestershire, UK).

His-elute buffer B: 50mM Tris-HCl pH8.0, 300mM NaCl, 1M imidazole.

1ml of metal-chelating cellulose (Bioline, London, UK) was loaded onto a column and washed twice with 5ml sterile Milli-Q water. The column was charged with 2ml 40mM $NiCl_2$ (Sigma, Gillingham, UK) and washed with Milli-Q water to remove the excess. The column was equilibrated with 2x 5ml his bind-buffer A. The cell pellet from a 100ml expressed culture was treated as mentioned above with no PEI treatment. The filtered supernatant was loaded onto the nickel charged cellulose and the flow-through collected. The column was then washed with 5ml of his-bind buffer A and bound protein eluted using a step gradient of the his-elute buffer B. 5ml steps of 7%, 14% and 40% his-elute buffer B were performed and 1ml fractions collected. A wash of 100% his-elute buffer B was carried out as a final step and the flow-through collected. A sample of each fraction was electrophoresed on a 12.5%(v/v) SDS PAGE gel, along with an aliquot of both the load and the flow-through to verify the presence of the correct-sized protein. In addition, an activity assay could also be performed to test for active protein in the fractions.

Heparin Column

Heparin-bind buffer: 20mM Tris-HCl pH8.0, 0.1mM EDTA, 0.1%(v/v) Tween-20, 25mM NaCl

Heparin-elution buffer: 20mM Tris-HCl pH8.0, 0.1mM EDTA, 0.1%(v/v) Tween-20, 2M NaCl.

A 5ml HiTrap™ Heparin HP column (GE Healthcare, Chalfont St. Giles, UK) was equilibrated with 5 column volumes (cv) of heparin-bind buffer. A protein sample, prepared in the appropriate buffer as per the protocol, was loaded onto the column at 1.5ml/min. The flow-through was collected, and the column washed with an additional 2cv of heparin-bind buffer. The proteins were eluted from the column using a gradient of 0.025M to 2M NaCl over 20cv at 1.5ml/min.

1.4ml fractions were collected. A sample of each fraction corresponding to a peak of A_{280} on the purification traces was electrophoresed on an SDS PAGE gel along with an aliquot of both the load and the flow-through to verify for the presence of the correct-size protein. In addition, an activity assay could also be performed to test for activity in the protein fractions.

Cation Exchange – SP Column

SP-Load Buffer: 25mM sodium phosphate (Fisher Scientific, Leicestershire, UK), 0.1mM EDTA, 25mM NaCl, pH6.8.

SP-elute buffer: 25mM sodium phosphate, 0.1mM EDTA, 2M NaCl.

5x 1ml HiTrap™ SP Hp column (GE Healthcare, Chalfont St. Giles, UK) were equilibrated with 5cv of SP load buffer. A protein sample, prepared in the SP load buffer, was loaded onto the column and the flow-through collected. The column was then washed with an additional 2cv of SP load buffer. Proteins were eluted from the column with a gradient of 0.025M to 2M NaCl over 20cv at 1ml/min and collected in 1.4ml fractions. A sample of each fraction, corresponding to an A_{280} peak on the purification trace, was electrophoresed on a 12.5%(v/v) SDS PAGE gel along with a sample of the load and the flow-through to verify the presence of the correct-size protein. An activity assay could also be performed to test for active proteins in the fractions.

Anion Exchange – Q Column

Q-bind buffer: 25mM Tris-HCl, 0.1mM EDTA, 0.1%(v/v) Tween-20, 25mM NaCl, pH8.0.

Q-elute buffer: 25mM Tris-HCl, 0.1mM EDTA, 0.1%(v/v) Tween-20, 2M NaCl, pH8.0.

A 5ml HiTrap™ Q HP column (GE Healthcare, Chalfont St. Giles, UK) was equilibrated with Q-bind buffer. A cell pellet prepared in the same buffer was loaded onto the column at 1.5ml/min and the flow-through collected. The column was washed with 2cv of Q-bind buffer and the proteins eluted using a gradient of 0.025M-2M NaCl over 20cv and 1.4ml fractions collected. A sample of each fraction corresponding to an A_{280} peak on the purification trace was electrophoresed on an SDS PAGE gel along with a sample of the load and the flow-through to verify the presence of the correct-size protein. An activity assay could also be performed to test for active proteins in the fractions.

Large-Scale His Purification

His-bind buffer C: 20mM Tris-HCl pH7.9, 500mM NaCl, 5mM Imidazole.

His-elute buffer D: 20mM Tris-HCl pH7.9, 500mM NaCl, 400mM imidazole.

A 14ml IDA Sepharose column was charged with Ni^{2+} by loading 2cv of 50mM $NiCl_2$. The column was then washed with sterile Milli-Q water, to remove the excess and equilibrated with his-bind buffer C. A cell pellet prepared in same buffer was loaded onto the column at 2ml/min and the flow-through retained. The his-tagged proteins were eluted from the column using a gradient of 5-400mM imidazole over 10cv at 2ml/min. 2ml fractions were collected. A sample of each fraction corresponding to an A_{280} peak on the purification trace was electrophoresed on a 12.5%(v/v) SDS PAGE gel along with a sample of the load and the flow-through to verify the presence of the correct sized protein. An activity assay could also be performed to test for active proteins in the fractions.

Ceramic Hydroxyapatite Column

CHT-bind buffer: 20mM potassium phosphate pH6.8, 25mM NaCl, 0.1%(v/v) Tween-20.

CHT-elute buffer: 400mM potassium phosphate pH6.8, 25mM NaCl, 0.1%(v/v) Tween-20.

A 19ml type I CHT column (BioRad, Hemel Hempstead, UK) was equilibrated with CHT-bind buffer. A prepared protein sample in CHT-bind buffer was loaded onto the column and the flow-through collected. The proteins were eluted using a phosphate gradient of 20 to 400mM over 10cv at 1.5ml/min. 2ml fractions were collected. A sample of each fraction corresponding to an A_{280} peak on the purification trace was electrophoresed on a 12.5%(v/v) SDS PAGE gel along with a sample of the load and the flow-through to verify the presence of the correct sized protein. An activity assay could also be performed to test for active proteins in the fractions.

Basic Enzyme Activity Assay

Samples and fractions were tested for RT activity using two steps; the first generated a cDNA strand from an RNA template and the second step involved a PCR to amplify the cDNA to allow its detection on an agarose gel.

<u>cDNA Synthesis Step</u>

The cDNA synthesis reaction contained 1x RT buffer (GeneSys Ltd, Camberley, Surrey), 0.5mM dNTPs, 15pmol MS2:3395_R primer (Appendix I), 20ng MS2 RNA (Roche, Welwyn Garden City, UK), 0.35µl of enzyme sample and the volume made to 20µl with nuclease-free water. Reactions were incubated at

45ºC for 30min to allow cDNA synthesis to occur, followed by 95ºC to denature the RT enzyme. This product could then be used as a template in a PCR.

cDNA Amplification

The PCR was set up with the final reaction containing 1x *Taq* master mix (GeneSys Ltd, Camberley, Surrey), 12.5pmol MS2:3231_F primer (Appendix I) and 12.5pmol MS2:3395_R primer, 0.5µl of the completed cDNA synthesis reaction and the final volume made up to 25µl with nuclease-free water. The reactions were cycled as follows:

94ºC	3min		
94ºC	10s	}	30 cycles
55ºC	10s		
72ºC	20s		
72ºC	7min		

5µl of the reactions were loaded on an agarose gel with the presence of the correct-size band of 164bp indicating RT activity.

Enzyme Storage

Enzyme storage buffer: 20mM Tris-HCl pH7.6, 0.2mM EDTA, 2mM DTT, 400mM NaCl, 0.02% Igepal, 50%(v/v) glycerol. All reagents were nuclease-free.

Once purified, the enzyme was concentrated using a Millipore Amicon Ultra-15 centrifugal filter (Fisher Scientific, Leicestershire, UK). This filter was also used to buffer exchange the enzyme into a 2x enzyme storage buffer with no glycerol.

Once concentrated to approximately 750µl in the 2x storage buffer, an equal volume of 100% glycerol was added. The protein could then be stored at either -20 or -70ºC.

Protein Concentration

2ml Protein Assay Dye Reagent Concentrate (BioRad, Hemel Hampstead, UK) was diluted with 7ml of Milli-Q water and 900µl dispensed into six 1cm path-length cuvettes. Bovine Serum Albumin (BSA) (BioRad, Hemel Hampstead, UK) was diluted to 100µg/ml stock, volumes of 0, 20, 50, 80 and 100µl were added to the cuvettes and each solution was made up to 100µl with Milli-Q water. The final cuvette had 1µl of the purified protein sample added and the volume made up to 100µl with Milli-Q water. The cuvettes were incubated at room temperature for 10min and then measurements taken at A_{595}. The measurements from the known BSA concentrations can be plotted onto a graph (Figure 4.2) allowing an estimation of the unknown concentration of the purified protein.

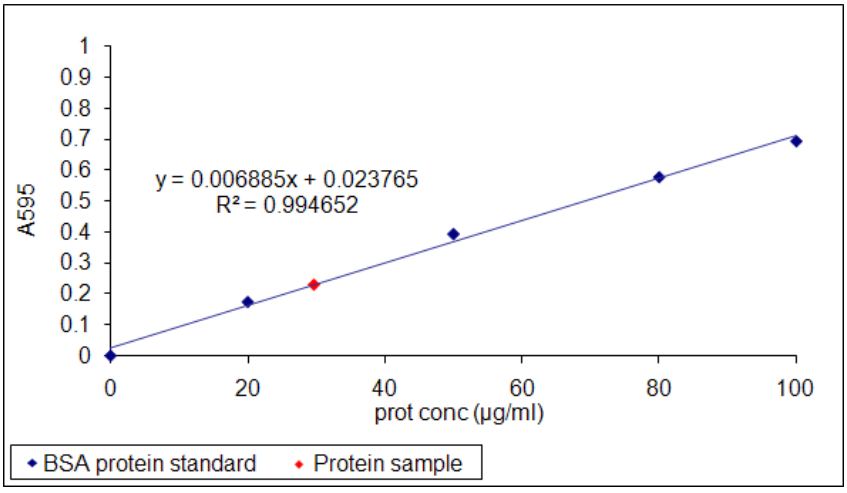

Figure 4.1: A standard curve showing the results from a typical assay to estimate protein concentrations. Known concentrations of BSA were measured at A_{595} and plotted on the graph (blue) allowing an estimation of the concentration of a protein sample (red).

4.3 – RESULTS

B. caldovelox IEP Purification

Heat Treatment

Large-scale expression of the IEP and the N-terminal his-tagged IEP from *B. caldovelox* was successful using TB medium with OD_{600} reaching approximately 6.6. No IEP was present in the insoluble fractions (Figure 4.3). Given that the IEP is from a bacterium with an optimum growth temperature of 70ºC, the cell extract was heat-treated as a clean-up step to remove some of the background *E. coli* proteins. The clarified cell lysate from the his-tagged protein was heated at 5min intervals at 50ºC and both soluble and insoluble fractions electrophoresed on a protein gel (Figure 4.3).

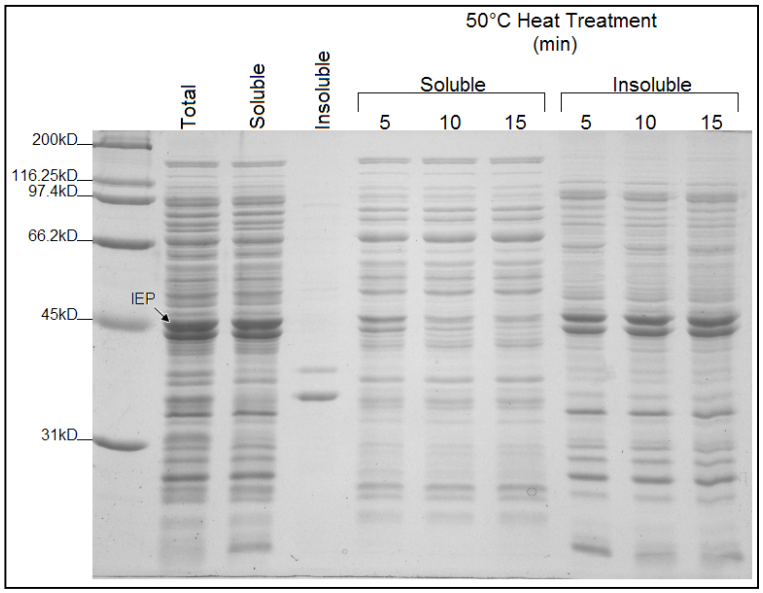

Figure 4.3: 12.5% SDS PAGE gel showing the total cell lysate fractions, the soluble and insoluble protein fractions, and the effect of various intervals of 50ºC on the N-terminal his-tagged IEP from *B. caldovelox*.

The protein gel showed that, even after 5min at 50ºC, most of the IEP had become insoluble. This experiment was repeated for the non-tagged IEP to ensure that the his-tag was not having an effect on the stability of the protein; however, the results were the same. Therefore heat treatment of the cell extract could not be used to reduce the background of *E. coli* proteins in the sample.

Nickel Column

The N-terminal his-tag on the IEP was created to enable an easy purification step allowing the his-tagged protein to bind to a Ni^{2+} charged column whilst the majority of the background proteins would come out in the flow-through. The cell pellet from a 250ml expression sample was prepared and the soluble fraction loaded onto a 1ml bench-top, metal-chelating column charged with Ni^{2+}. The flow-through was collected along with 5x 1ml fractions from the step gradient of 5%, 7%, 14%, 40% and 100% his-elute buffer. When the fractions, along with the flow-through and initial load, were loaded on a protein gel, it became apparent that no proteins had bound to the column and that the his-tagged IEP had come out in the flow-through (Figure 4.4).

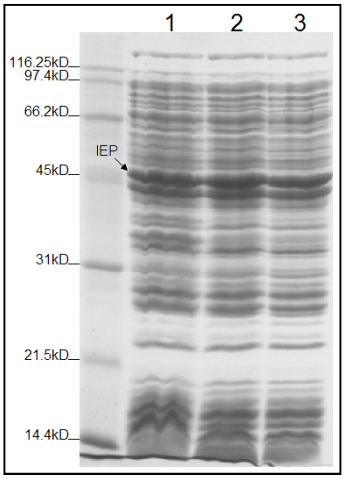

Figure 4.4: A 12.5% SDS PAGE gel showing lane 1: the total cell fraction, lane 2: the filtered soluble fraction loaded onto the column, and lane 3: the flow-through from the Ni^{2+} charged column.

This experiment was also repeated with a cell pellet from 100ml of expressed culture in case overloading of the column was preventing the protein from binding; however, still no protein bound to the column.

No further purification steps were attempted on the his-tagged IEP; instead, all further purification steps were performed on the non-tagged protein.

SP Cation Exchange Column

The pI value of the protein was calculated from the amino acid composition of the protein and determined to be 10.14. Therefore at a pH6.8 the protein would carry a net positive charge, allowing it to bind to an SP cation exchange column. The cell pellet from 100ml of expressed culture was prepared and loaded onto a 1ml SP column and eluted using a 0.025M-2M NaCl gradient. However, the protein trace revealed no protein to elute from the column as all proteins had come through in the initial flow-through.

Anion Exchange Q column

Since the protein failed to bind to a cation exchange column, it was decided to use an anion exchange column as a purification step. At pH8.0 the IEP would still have a net positive charge and it was expected that this protein would come out in the flow-through purer than the load sample since the majority of the *E. coli* background proteins would bind to the column. The pellet from a 250ml expressed culture was prepared as before and loaded onto a 5ml HiTrap™ Q column with a gradient of 0.025-2M NaCl performed over 20cv. Although the IEP was expected to be found in the flow-through, fractions from the protein peaks seen on the purification trace (Figure 4.5) were also electrophoresed on a protein gel (Figure 4.6). Unexpectedly, it was found that all the IEP had bound

to the column and eluted at a high NaCl concentration of approximately 1.3M, suggesting tight binding of the protein to the column.

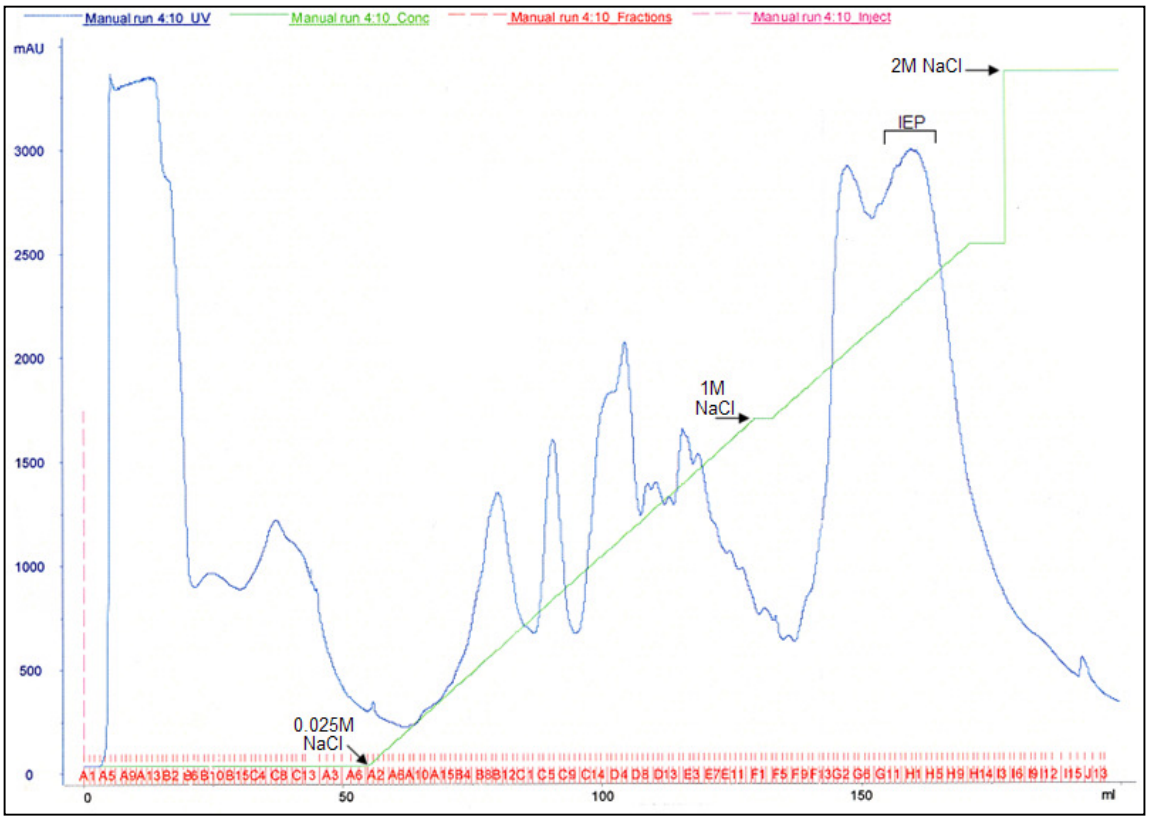

Figure 4.5: The purification profile of the *B. caldovelox* IEP loaded on and eluted from a 5ml HiTrap™ Q column. The protein eluted within the peak marked IEP at approximately 1.3M NaCl.

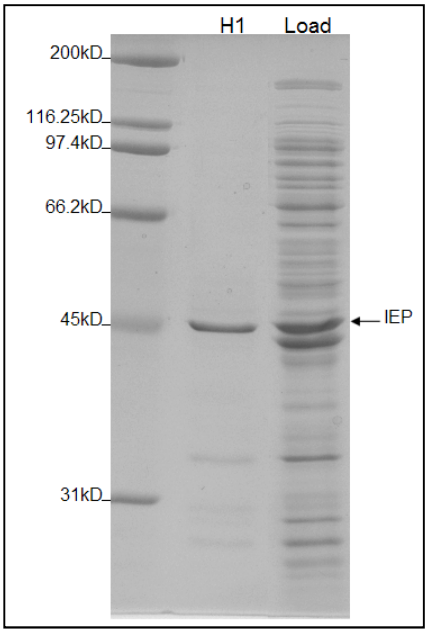

Figure 4.6: 12.5%(v/v) SDS PAGE gel showing a 1µl aliquot of the H1 fraction from the HiTrap™ Q column elution.

Scanning fraction H1 over an absorbance range of A_{260}-A_{280} revealed a large peak at A_{260} indicating a high degree of contaminating nucleic acids within the IEP elution fractions. Although the IEP was relatively free of protein contaminations, the association with nucleic acids prevented the use of the Q column as a purification step.

Heparin Column

It was decided to use a different approach to purification and to take advantage of the nucleic acid binding properties of the protein. A 5ml HiTrap™ Heparin column was used to allow nucleic acid binding proteins to bind and the remaining *E. coli* proteins to wash straight through. However, when a 250ml expressed culture was prepared as before and loaded onto the column, the IEP did not bind to the column.

It was hypothesised that the IEP has a higher affinity for nucleic acids than for the heparin column. This association with the nucleic acids might alter its charge, explaining why it would not bind to the cation exchange column and why, when bound to the anion exchange column, it eluted at a high NaCl concentration with the nucleic acid fraction. Therefore, in order to purify this IEP it would seem necessary to remove the nucleic acids prior to any purification steps.

PEI Precipitation

The PEI precipitation had to be optimised for every new expression culture to ensure that the appropriate level of PEI is added to remove all nucleic acids and, if necessary, the correct concentration of NaCl is used to ensure all the IEP remained soluble. In order to remove nucleic acids, a sample from a large expressed culture was removed and the pellet re-suspended in the equivalent of 1ml/0.1g of cell pellet in the heparin-bind buffer. After sonication, the total protein fraction was exposed to varying levels of PEI and electrophoresed on an agarose gel (Figure 4.7a), to detect removal of nucleic acids, and a protein gel (Figure 4.7b) to ensure that the IEP remained in solution.

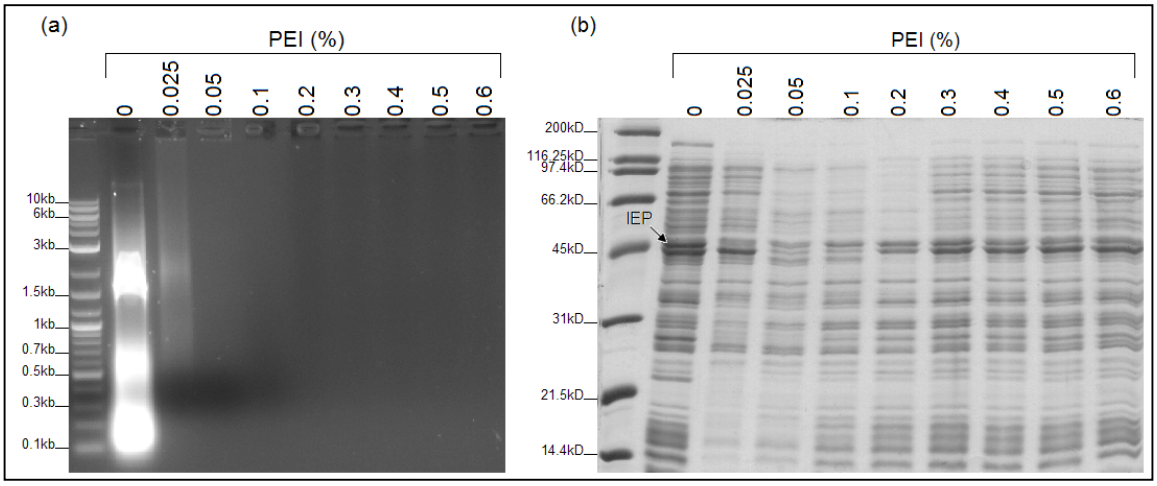

Figure 4.7:

a) A 1%(w/v) agarose gel showing the removal of nucleic acids from the protein sample with increasing levels of PEI.

(b) A 12.5%(v/v) SDS PAGE gel showing the effect of increasing levels of PEI on the soluble IEP fraction.

The agarose gel in Figure 4.7a showed that 0.05% PEI was necessary to remove the contaminating nucleic acids. However, at this level of PEI the protein gel revealed that the IEP had become insoluble. The protein became insoluble at the lowest levels of PEI; however, at higher PEI concentrations than necessary to remove all nucleic acids the protein reappeared in the soluble fractions. Since high levels of PEI can interfere with subsequent purification steps it was necessary to repeat the experiment with increasing concentrations of NaCl to identify a level at which the IEP would remain in solution at the minimum level of PEI required to remove all nucleic acids. 0.05% PEI was used with varying concentrations of NaCl and the soluble fraction electrophoresed on a protein gel (Figure 4.8).

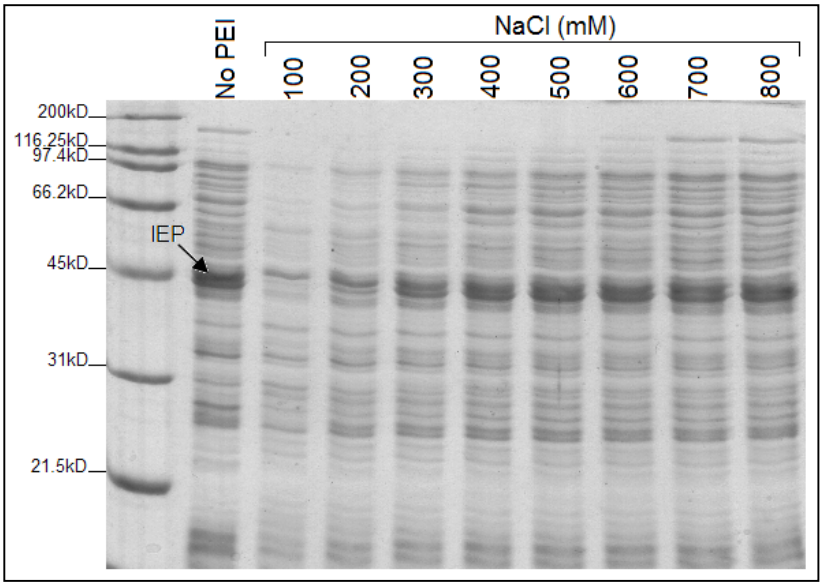

Figure 4.8: A 12.5%(v/v) SDS PAGE gel showing the effects on the soluble IEP at 0.05% PEI with varying concentrations of NaCl.

Varying the concentration of NaCl with 0.05% PEI showed that 500mM NaCl precipitated the nucleic acids but to allow the majority of the IEP to remain soluble.

Heparin Column – After PEI Precipitation

The cell pellet from 250ml of an expression culture was re-suspended in a heparin-bind buffer containing 500mM NaCl, sonicated and treated with 0.05% PEI. After incubation at 4ºC for 1h the sample was centrifuged and the supernatant filtered and loaded onto the heparin column and the flow-through collected. To ensure that the IEP remained in solution after PEI precipitation, and also to see if it had bound to the column, fractions from each stage were electrophoresed on a protein gel (Figure 4.9).

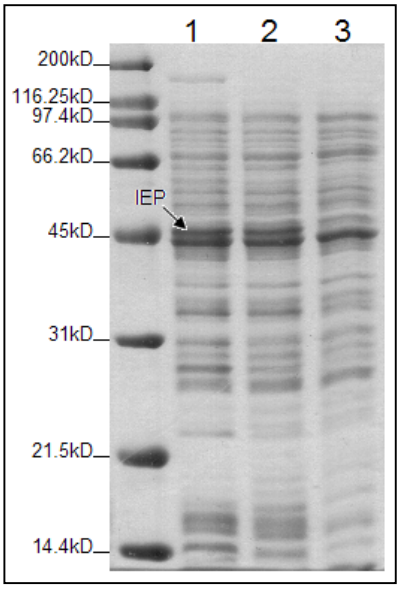

Figure 4.9: 12.5%(v/v) SDS PAGE gel showing: Lane 1: the soluble protein before PEI precipitation, lane 2: the soluble fraction after PEI precipitation and filtering and lane 3: the flow-through from the heparin column.

Once it was verified that the IEP had bound to the heparin column proteins were eluted using a gradient of 0.5-2M NaCl. Fractions corresponding to A_{280} peaks on the purification trace (Figure 4.10) were electrophoresed on a protein gel to reveal where the protein had eluted. It was shown that the fractions of the A_{280} peak corresponding to approximately 1.6M NaCl contained a protein at the expected size for the IEP (Figure 4.11a). The peak protein fraction D12 was also electrophoresed on an agarose gel (Figure 4.11b) which confirmed the absence of nucleic acids from the sample.

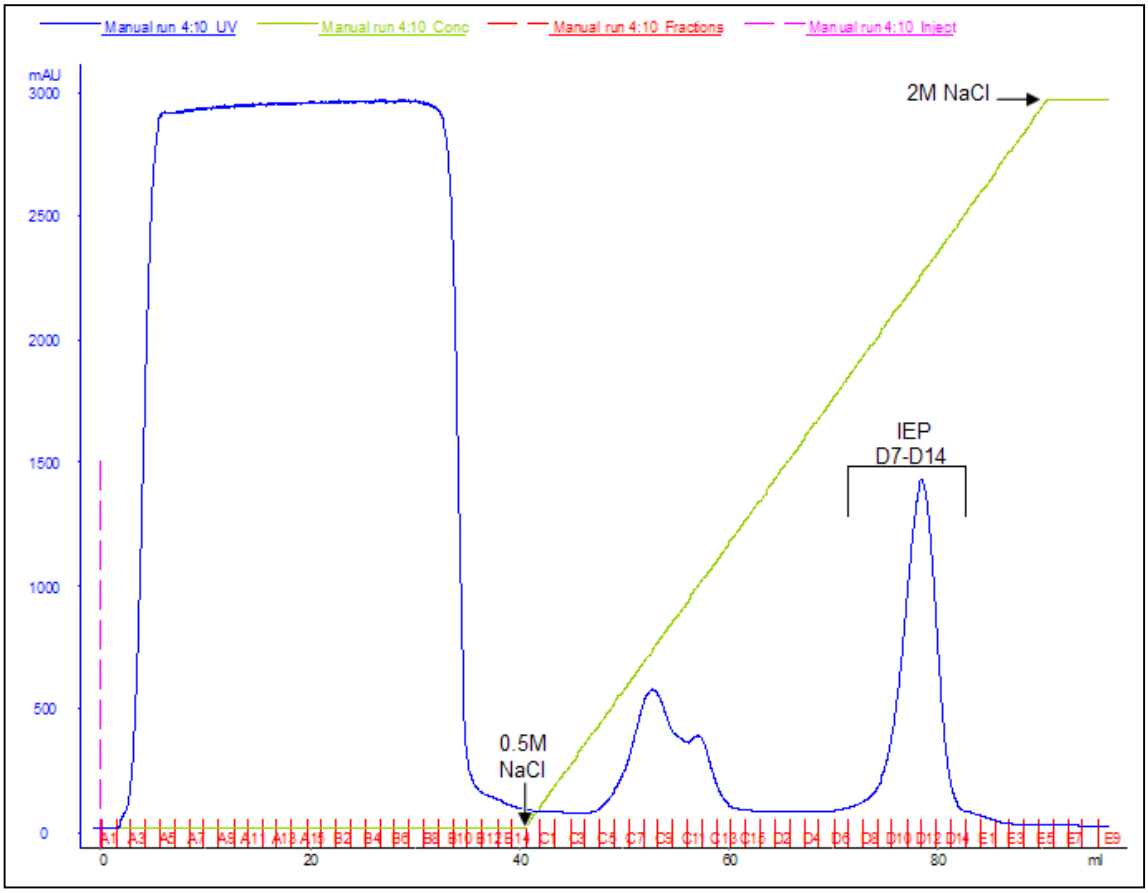

Figure 4.10: A purification profile of the *B. caldovelox* IEP loaded onto a heparin column after precipitating with PEI to remove nucleic acids. The IEP was located in the final protein peak at approximately 1.6M NaCl.

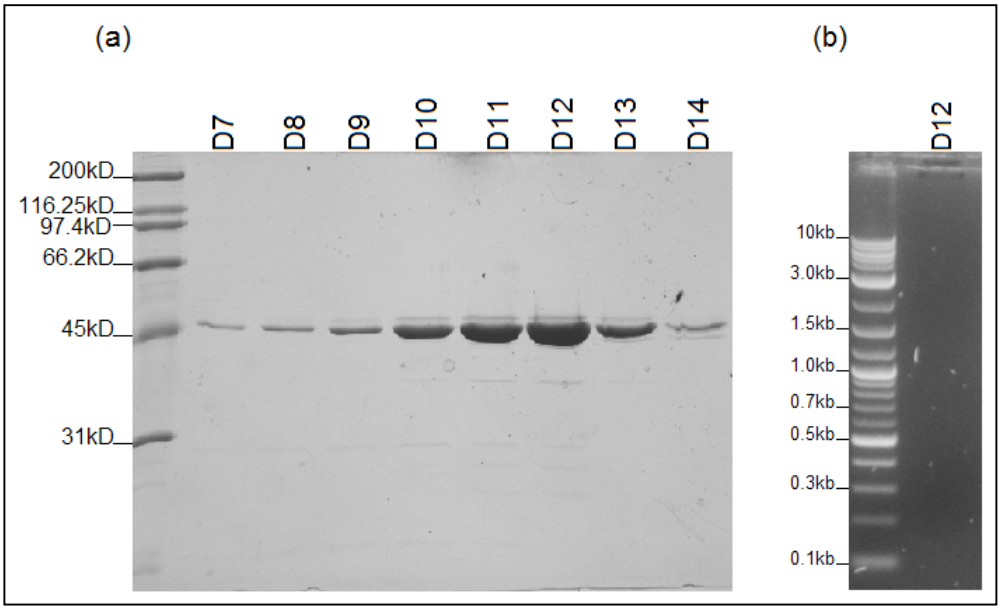

Figure 4.11:

a) A 12.5% SDS PAGE gel showing 5µl of the fractions from the IEP elution peak.

b) A 1.5%(w/v) agarose gel with fraction D12 confirming the absence of nucleic acids from the sample.

To confirm that these fractions did contain the IEP, it was tested for RT activity in a two step assay. A dilution equivalent to 0.0625µl of the D12 fraction was used in a cDNA synthesis step with MS2 RNA and MS2:3193_R primer (Appendix I). A no-enzyme control (NEC) was run in parallel to ensure that any DNA amplified in the second step using PCR was only due to the RT activity of the fraction. Loading an aliquot of each PCR on an agarose gel confirmed the presence of the correct size band of 718bp from using fraction D12 compared to no band on the NEC (Figure 4.12). This indicated the presence of RT activity and therefore the correct IEP within this protein peak.

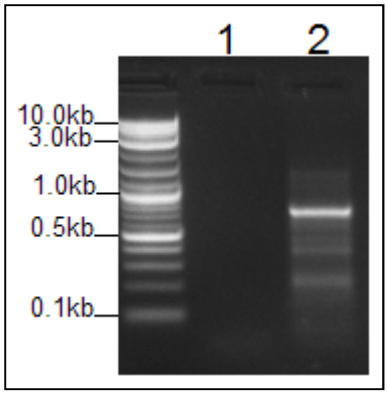

Figure 4.12: A 1.5%(w/v) agarose gel showing the result of a PCR using the cDNA synthesis reaction as a template. Lane 1: no RT enzyme control, and lane 2: When a 0.0625µl aliquot of the D12 fraction was used as the RT enzyme.

SP Column – After PEI Precipitation and Heparin Purification

Fractions D9-D13 were pooled and dialysed into the cation-bind buffer. This protein sample was loaded onto 5x 1ml HiTrap™ SP columns and eluted using a 20cv gradient of 0.025-2M NaCl. This time the IEP did bind to the column and eluted at approximately 1.6M NaCl. Therefore the nucleic acids were previously interfering with the enzyme preventing its binding to the cation exchange column. However, the cation exchange column lowered the yield of the enzyme without any increase in purity. It was therefore decided that the heparin column alone was enough to purify the IEP to allow its characterisation as an RT enzyme.

The heparin purification step was repeated on a 250ml protein expressed culture and the appropriate fractions concentrated and buffer exchanged into a storage buffer. This gave approximately 1.5ml of very pure IEP (Figure 4.13) with a protein concentration of 6.4mg/ml estimated using a standard curve of known concentrations of BSA.

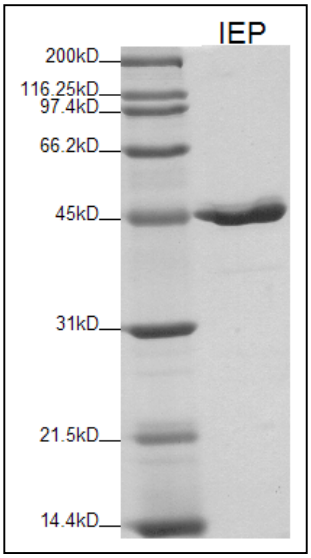

Figure 4.13: A 12.5% SDS PAGE gel showing 1µl of the heparin purified IEP from *B. caldovelox*.

This protein band was excised from the gel and sent off for mass spectrometry analysis which confirmed the identity of the protein as being a correct match to the GK1355-IEP as expected.

IEP-Sac7d and Sac7d-IEP Fusion Protein Purification

The two different Sac7d fusion proteins were successfully expressed on large-scale using TB media, with the cultures reaching an OD_{600} between 6.5-6.6.

Heat Treatment

The soluble protein fractions from the Sac7d fusion protein expression cultures were heated to 50ºC for 5min. After centrifugation, the soluble fraction was electrophoresed on a protein gel (Figure 4.14) which showed that the protein, like the wild-type IEP, had become insoluble and therefore heat treatment was not a method that could be used to reduce background *E. coli* proteins.

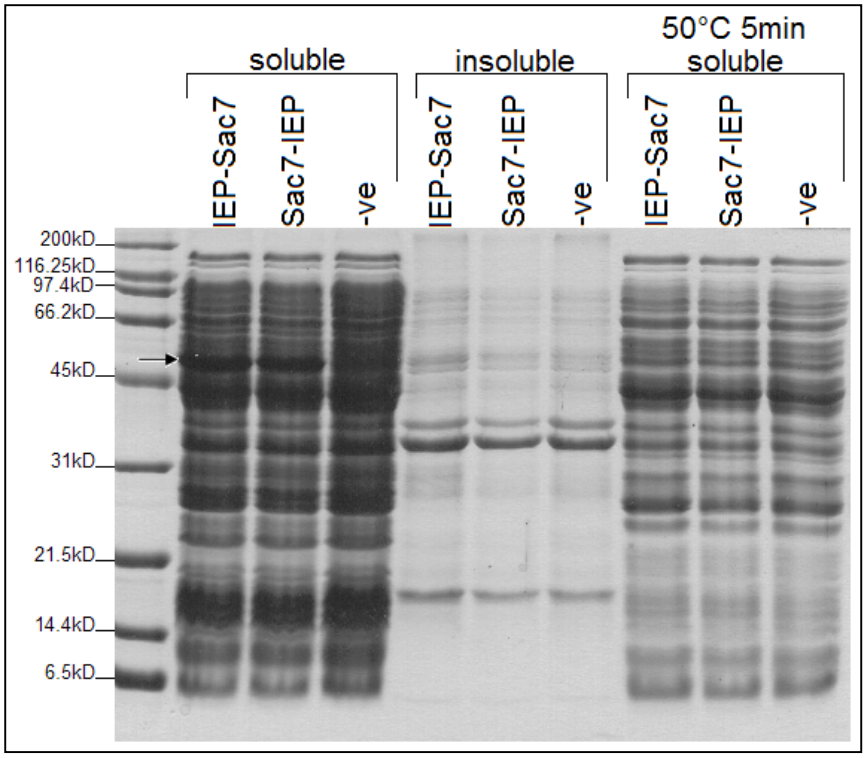

Figure 4.14: A 12.5%(v/v) SDS PAGE gel showing the soluble and insoluble protein fraction from the large-scale expression and the effect on the soluble IEP of incubation for 5min at 50ºC. The arrow indicates soluble IEP-Sac7d that is not present in the negative-control lane (-ve).

IEP-Sac7d

PEI Precipitation

IEP-Sac7d was over-expressed on a large-scale, a sample removed and the cell pellet from the sample re-suspended in heparin-bind buffer equivalent to 1ml/0.1g of cells. The cell sample was sonicated and the total protein fraction tested, as before, with varying levels of PEI to remove nucleic acids. After incubating for 1h at 4ºC the samples were centrifuged and electrophoresed on an agarose gel (Figure 4.15a) to see the concentration required to remove all the nucleic acids from the sample, and onto a protein gel (Figure 4.15b) to see the effect of the PEI on the soluble IEP-Sac7d protein.

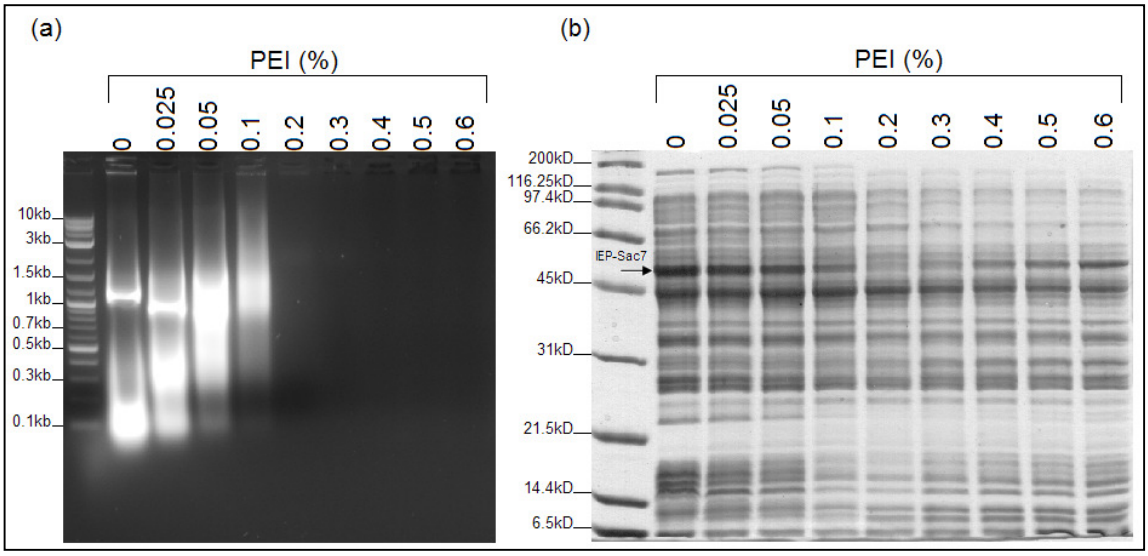

Figure 4.14:

a) A 1%(w/v) agarose gel showing the effect on the nucleic acids within the protein sample with increasing levels of PEI.

b) A 12.5%(v/v) SDS PAGE gel showing the effect of the increasing PEI concentration on the solubility of the IEP-Sac7d.

The agarose gel showed that 0.2% PEI was required to remove the nucleic acids from this sample; however, at this level, all the IEP-Sac7d protein was precipitated. Therefore it was necessary repeat the experiment with varying concentrations of NaCl and the soluble fraction was electrophoresed on a protein gel (Figure 4.16).

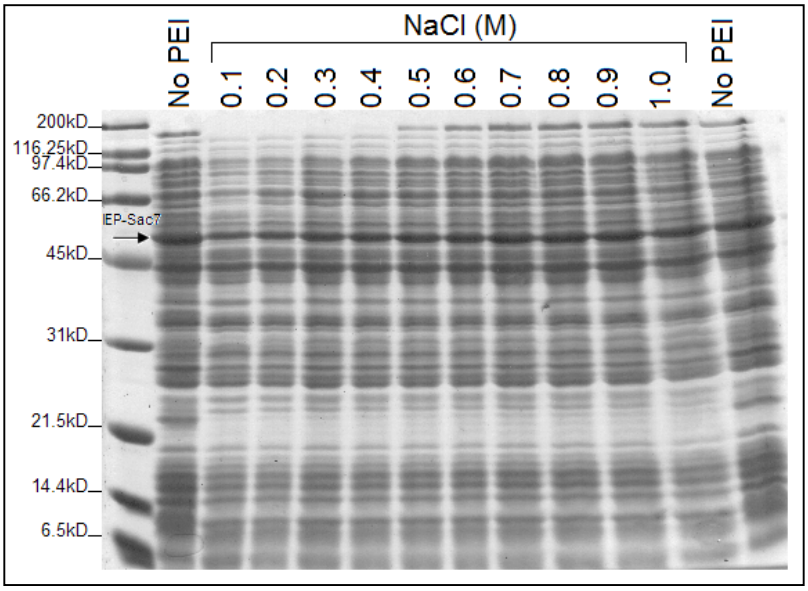

Figure 4.16: A 12.5% SDS PAGE gel showing the effect on the soluble IEP-Sac7d protein at 0.2% PEI with increasing concentration of NaCl.

Even at 0.1M NaCl, the majority of the IEP remained soluble. However, in order to obtain maximum yield, it was decided that 0.5M NaCl could be used.

Heparin Column – After PEI Precipitation

PEI precipitation was scaled up to be used with a cell pellet from a 250ml expressed culture. The pellet from this culture was re-suspended in heparin-bind buffer (0.5M NaCl) equivalent to 1ml/0.1g of cells. After sonication the protein sample was treated with 0.2% PEI, incubated at 4ºC for 1h and then centrifuged. The soluble fraction was filtered and loaded onto a 5ml heparin column and the flow-through collected. Samples of the protein fraction pre-PEI precipitation, after PEI precipitation, and the flow-through from the column were electrophoresed on a protein gel (Figure 4.17).

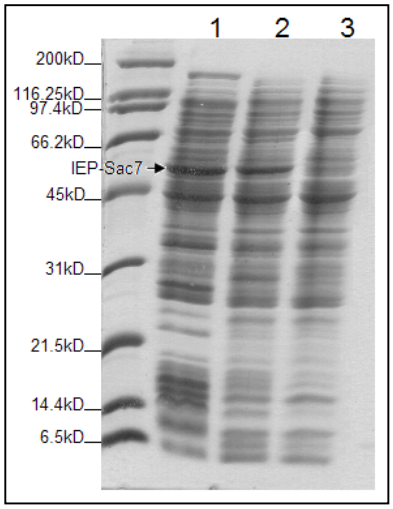

Figure 4.17: A 12.5% (v/v) SDS PAGE gel showing: lane 1: the total protein fraction before PEI precipitation, lane 2: the filtered soluble protein fraction after PEI precipitation, and lane 3: the flow-through from the heparin column.

After precipitation with PEI, all the IEP-Sac7d protein remained soluble and bound to the heparin column. Proteins were eluted from the column using a gradient of 0.5-2M NaCl over 20cv and 1.4ml fractions collected which revealed a purification trace very similar to that from the IEP with no Sac7d domain (Figure 4.18).

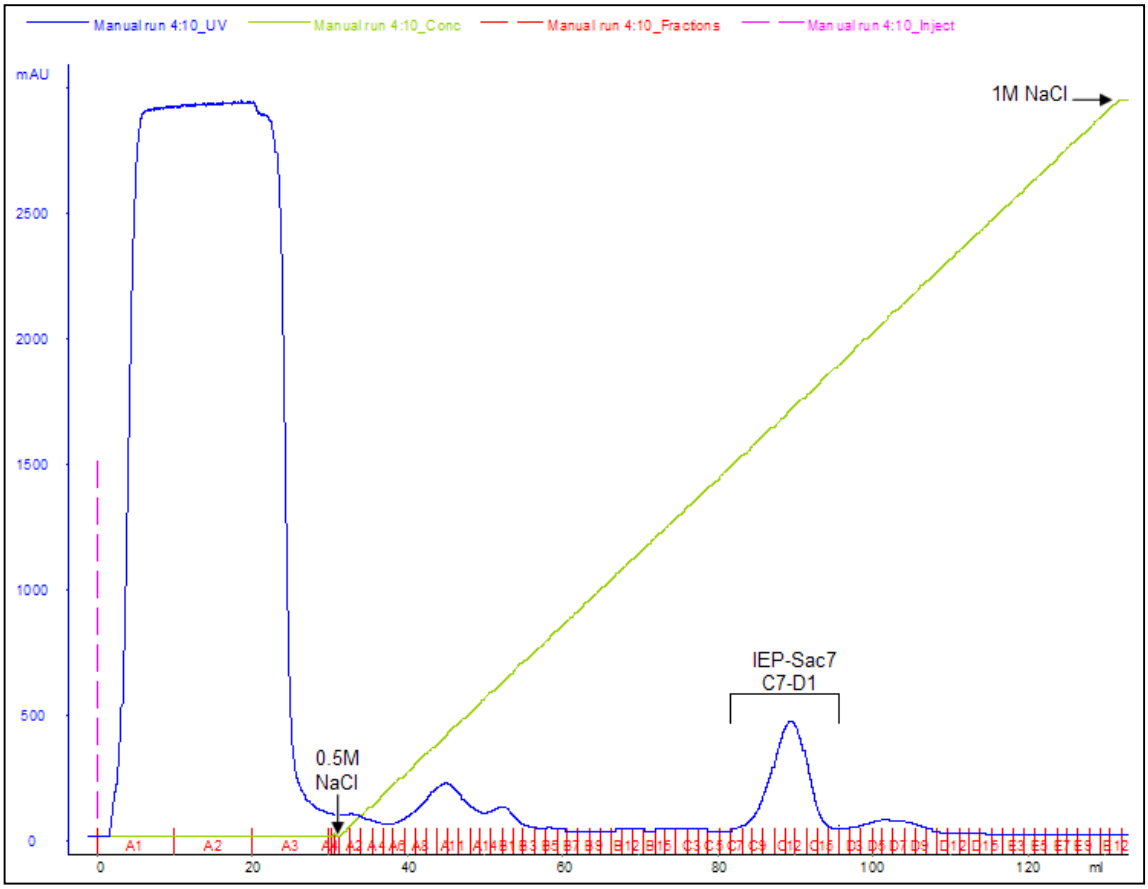

Figure 4.18: The purification trace of the load and elution of the IEP-Sac7d protein from a heparin column.

15µl aliquots of the fractions from the A_{280} protein peak as marked on the trace, C7-D1, were electrophoresed on a protein gel which showed pure protein of the correct size for Sac7d-IEP (Figure 4.19). This peak corresponded to an NaCl concentration of approximately 1.35M.

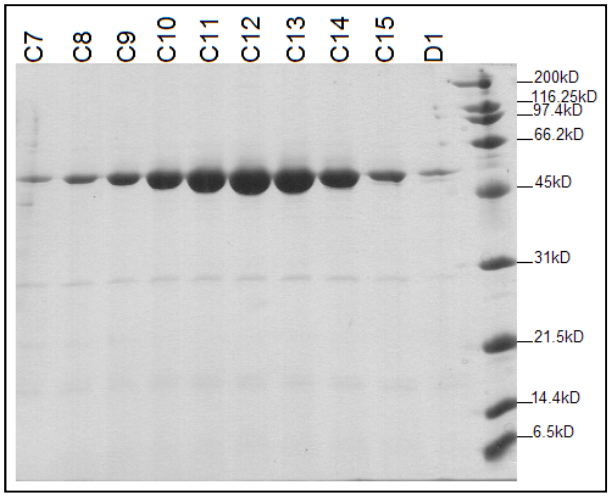

Figure 4.19: A 12.5% (v/v) SDS PAGE gel showing 15µl of the fractions across the peak at 1.25M NaCl M expected to contain the IEP-Sac7d.

Fractions C8-C14 were pooled, concentrated and the 1.5ml sample stored in enzyme storage buffer. To analyse the purity of the sample, 1µl was electrophoresed on an SDS PAGE gel (Figure 4.20) and, by using a standard curve, the protein concentration was estimated at 3mg/ml.

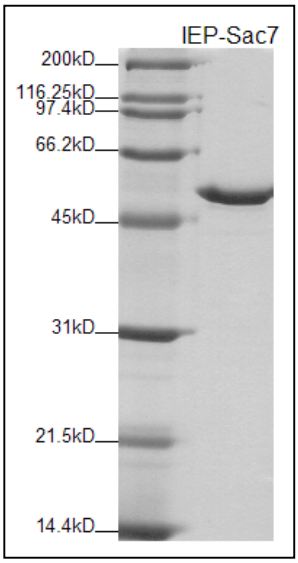

Figure 4.20: 12.5% SDS PAGE gel showing 1µl of the heparin purified IEP-Sac7d fusion protein.

Sac7d-IEP

PEI Precipitation

The Sac7d-IEP fusion protein was over-expressed on a large-scale and a sample removed for titration with PEI for nucleic acid removal. A cell pellet was re-suspended in heparin-bind buffer, equivalent to 1ml/0.1g of cells, and sonicated. Aliquots of the sonicated sample were subjected to varying concentrations of PEI. After incubation at 4ºC the samples were centrifuged and the soluble fraction loaded on to an agarose gel (Figure 4.21a) to identify the required level of PEI to remove all nucleic acids, and onto a protein gel to monitor the solubility of the Sac7d-IEP (Figure 4.21b).

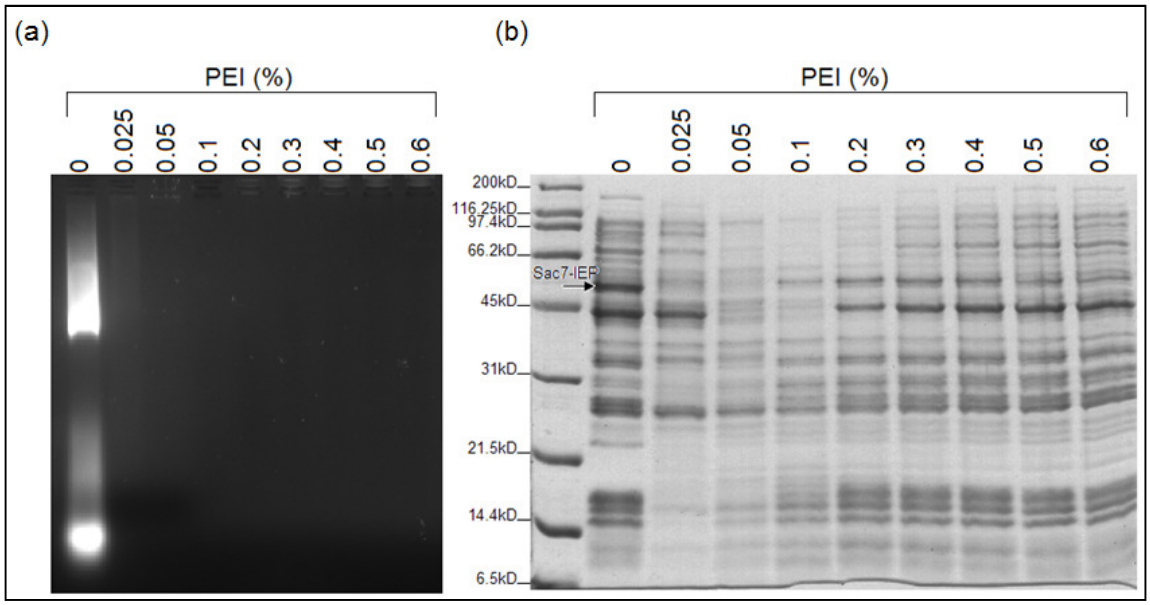

Figure 4.21:

a) A 1%(w/v) agarose gel showing the removal of nucleic acids from the protein sample with increasing concentrations of PEI.

b) A 12.5% SDS PAGE gel showing the effect of PEI on the solubility of the Sac7d-IEP.

The agarose gel showed that 0.05% PEI was required to remove all the nucleic acids from this sample; however, at this level all the Sac7d-IEP had become

insoluble. The experiment was repeated with 0.05% PEI and varying the concentration of NaCl. 5µl of the soluble fraction from this experiment was electrophoresed on a protein gel (Figure 4.22) to monitor the concentration of NaCl required to allow the Sac7d-IEP to remain in solution.

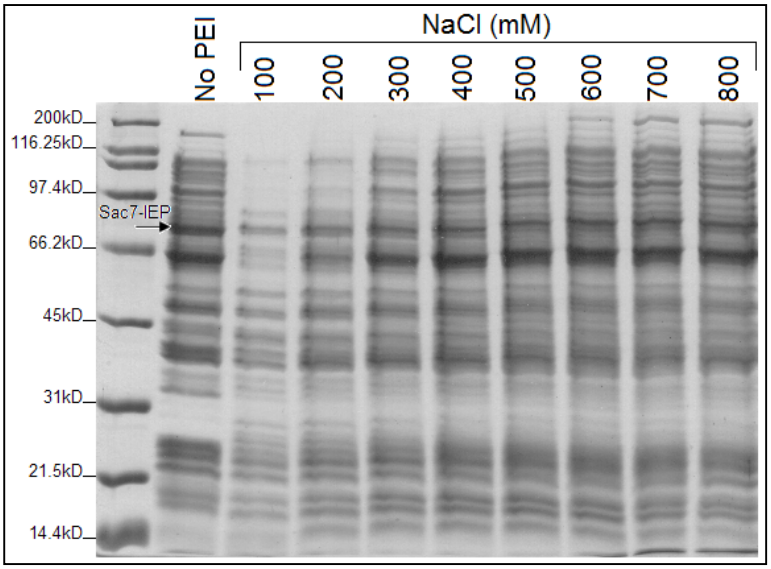

Figure 4.22: A 12.5% SDS PAGE gel showing the effect on the solubility of Sac7d-IEP when exposed to 0.05% PEI with varying concentrations of NaCl.

The protein gel showed that 500mM NaCl should allow the Sac7d-IEP to remain in solution when exposed to 0.05% PEI.

Heparin Column – After PEI Precipitation

The cell pellet from a 250ml expressed culture was re-suspended in heparin-bind buffer (500mM NaCl) equivalent to 1ml/0.1g of cells. The re-suspended pellet was sonicated and 0.05% PEI added and incubated as before. The sample was then centrifuged and the supernatant filtered and loaded onto a 5ml heparin column. Proteins were eluted from the heparin column using a gradient of 0.5-2M NaCl over 20cv, and 1.4ml fractions were collected. The purification profile (Figure 4.23) was similar to that of the IEP and the IEP-Sac7d with the

protein expected to be found within the A_{280} protein peak that corresponded to approximately 1.2M NaCl.

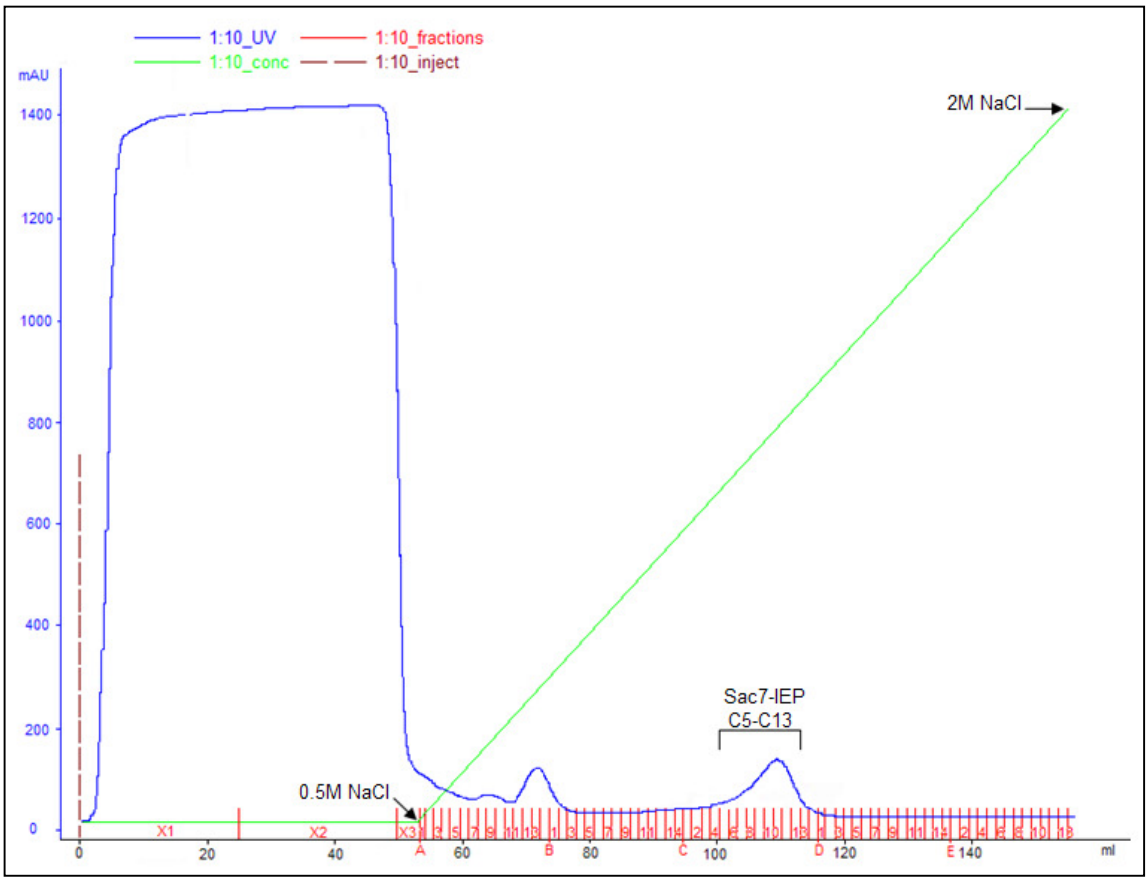

Figure 4.23: Purification profile showing the load and the elution of the Sac7d-IEP expression sample using a heparin column.

An SDS PAGE gel was loaded with 5µl samples of the protein fraction before PEI precipitation, the column load, the column flow-through and 15µl of the fractions across the 1.2M NaCl protein peak (Figure 4.24).

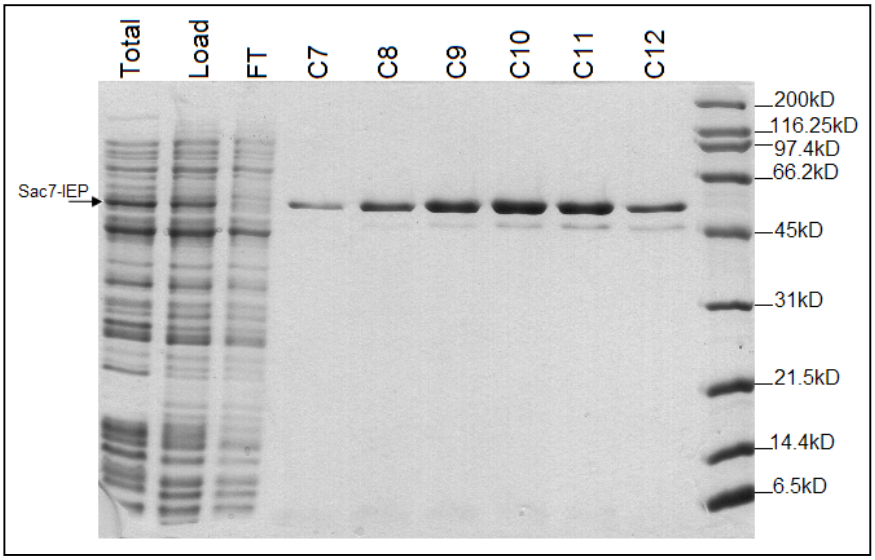

Figure 4.24: A 12.5%(v/v) SDS PAGE gel showing the total protein fraction before PEI precipitation, the PEI precipitated and filtered column load, the column flow-through and samples from the expected Sac7d-IEP peak.

Although some of the Sac7d-IEP had precipitated out along with the nucleic acids, some still remained and bound to the column; none of the protein came out in the flow-through. The fractions electrophoresed on the protein gel did contain the correct-size protein from Sac7d-IEP. However, there was the presence of low levels, approximately <10% of the entire sample, of an additional protein at approximately the same size as would be expected for the wild-type IEP.

Fractions C7-C12 were pooled and concentrated to a final volume of 1ml. 1µl of this sample was electrophoresed on a protein gel to analyse its purity (Figure 4.25) and the concentration was estimated, based on a BSA concentration standard curve, at 4.9mg/ml.

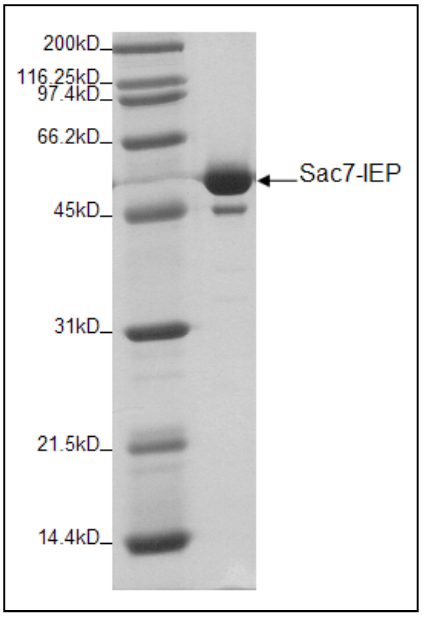

Figure 4.25: A 12.5%(v/v) SDS PAGE gel with 1µl of heparin purified Sac7d-IEP.

SP Column – After PEI Precipitation and Heparin Column

It was decided that an additional purification step would be required to remove the impurity that could be seen on the protein gel. The partially purified protein was dialysed into cation-bind buffer and loaded onto a 1ml SP column. The protein was eluted from the column with a gradient of 0.025-2M NaCl over 20cv and 1ml fractions were collected. Samples from all the protein peaks observed on the trace (Figure 4.26) were electrophoresed on a protein gel and it was found that the protein eluted in the major protein peak at 1.4M NaCl (Figure 4.27).

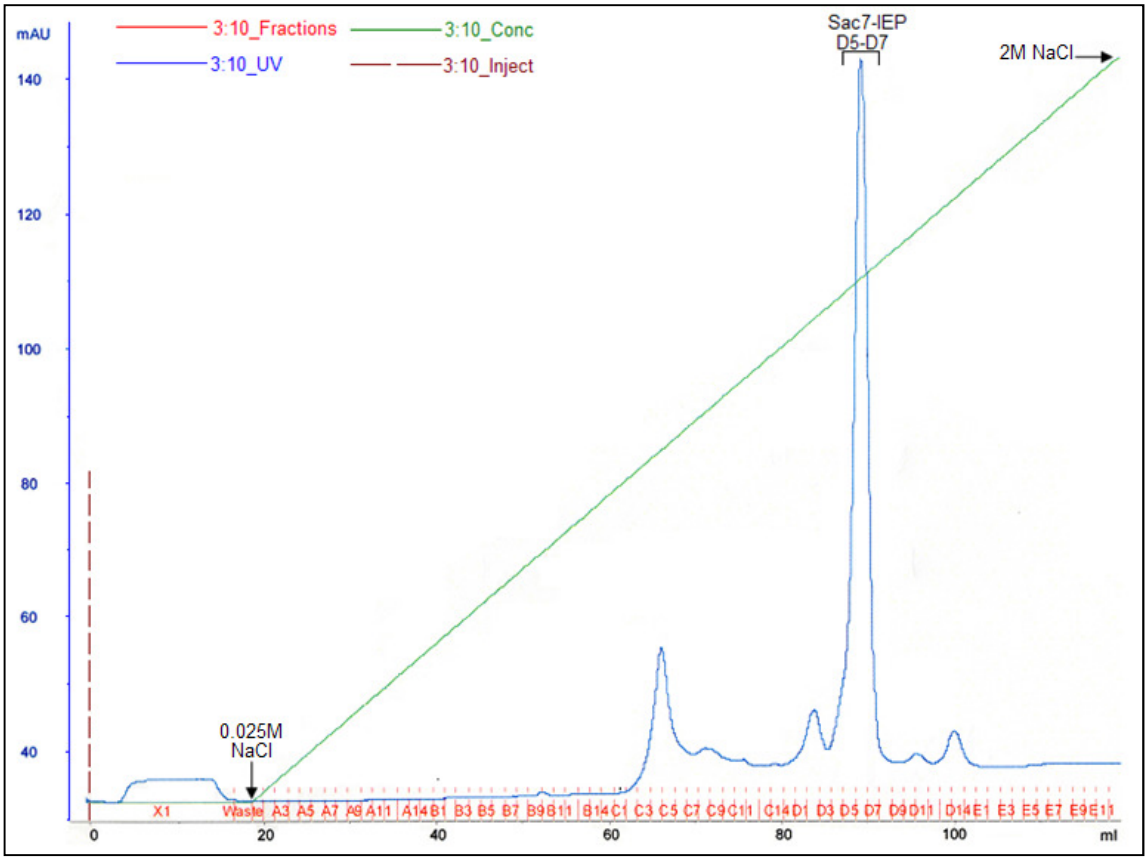

Figure 4.26: The purification trace of the heparin purified Sac7d-IEP loaded and eluted from a 1ml SP column.

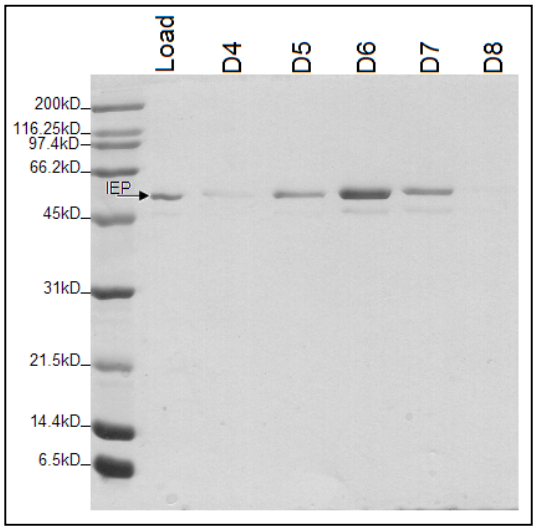

Figure 4.27: A 12.5% SDS PAGE gel showing 15µl samples of the fractions from the Sac7d-IEP protein peak at 1.4M NaCl.

The protein gel showed that the majority of the protein was present in fractions D5, D6 and D7 although the IEP had still co-purified with a contaminating protein. These fractions were pooled, concentrated and stored in enzyme storage buffer and 1µl electrophoresed on a protein gel (Figure 4.28).

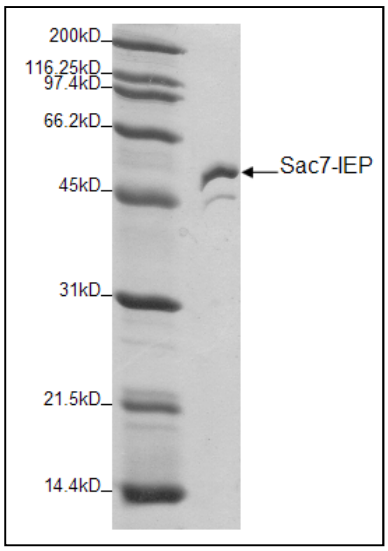

Figure 4.28: A 12.5%(v/v) SDS PAGE gel showing 1µl of the heparin and SP column purified Sac7d-IEP

The concentration of the Sac7d-IEP was estimated at 0.8mg/ml however, this was a combined concentration of both the Sac7d-IEP and the contaminant.

The co-purifying protein contaminant still seen after the two columns was not present in the purification of the IEP-Sac7d protein. This protein ran at the same size as would be expected for the wild-type IEP, and also has similar purification properties. It was therefore expected that this product was created due to the Sac7d linker being vulnerable to proteolytic cleavage. The purification was repeated using the protease inhibitor cOmplete, Mini, EDTA-free protease inhibitor cocktail (Roche, Welwyn Garden City, UK); however, the degradation product was still present suggesting that the cleavage was

occurring before cell lysis. It was therefore decided to remove the GTV linker allowing the two proteins to be fused together with no joining linker.

Linker Removal

New primers were designed (Appendix I) to allow the fusion of the two domains together into one ORF. Overlap extension PCR was successful producing the expected size fragment of 1461bp (Figure 4.29).

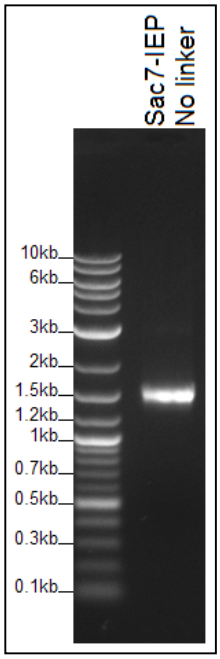

Figure 4.29: A 1% (w/v) agarose gel showing the final overlap extension PCR product fusing the gene products of *B. caldovelox* IEP and *S. acidocaldarius* Sac7d.

This overlap extension fragment was digested with *Nde* I and *Xho* I, ligated into pET24a(+) and transformed into electrocompetent *E. coli* KRX (pRARE2). Positive colonies were detected using a screening PCR with T7 promoter and terminator primers. Positive colonies, along with one negative-control, were grown in 10ml LB at 37℃ until an OD_{600} of 0.4 was reached. The temperature

was then reduced to 15℃ and protein expression induced according to the KRX protocol. After 20h of protein induction the cultures were centrifuged and the pellets re-suspended in 750µl cell lysis buffer D. The samples were then sonicated and 5µl of the soluble and insoluble fractions electrophoresed on a protein gel (Figure 4.30).

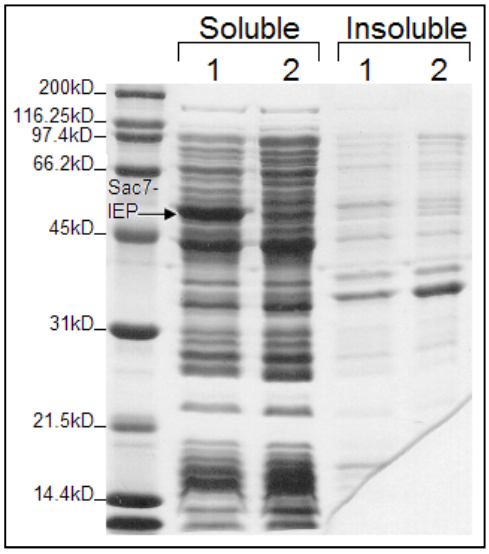

Figure 4.30: A 12.5% SDS PAGE gel showing the soluble and insoluble fractions from the over-expression of Sac7d-IEP (lanes 1) and a negative-control containing just pET24a(+) with no insert (lanes 2). The arrow marks the expected Sac7d-IEP.

The protein gel showed the appearance of a protein band, at the expected size of 56kD for Sac7d-IEP that was not present in the negative-control lane. This protein was 100% soluble and was then expressed on large-scale to allow purification.

PEI Precipitation

A sample from the expressed Sac7d-IEP culture was removed and re-suspended in heparin-bind buffer (0.5M NaCl) equivalent to 1ml/0.1g of cells. The sample was then centrifuged and exposed to varying concentrations of PEI

to precipitate the nucleic acids. These samples were then centrifuged and the soluble fraction loaded on to an agarose gel to verify the concentration required for removal of all the nucleic acids (Figure 4.31a), and onto a protein gel to analyse whether the Sac7d-IEP was precipitated with the required level of PEI (Figure 4.31b).

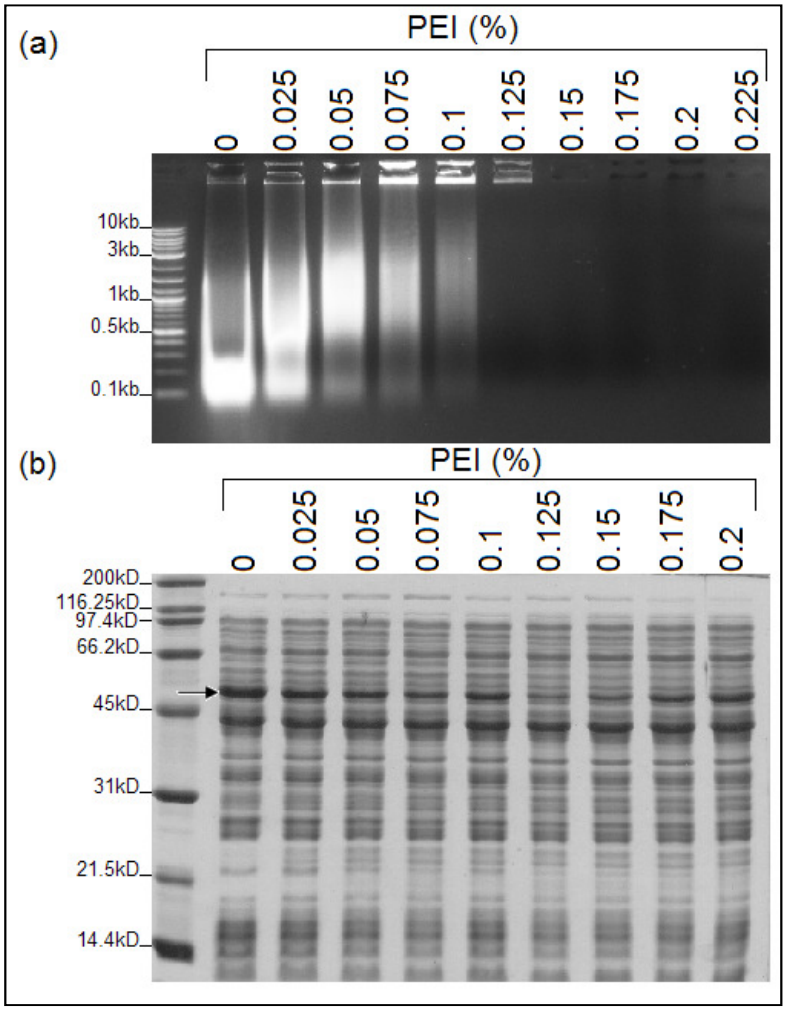

Figure 4.31:

a) A 1% (w/v) agarose gel showing the precipitation of nucleic acids from the sonicated protein sample with increasing concentrations of PEI.

b) A 12.5%(v/v) SDS PAGE gel showing the effect of increasing PEI on the solubility of Sac7d-IEP (marked with an arrow) in 500mM NaCl.

Heparin Column - After PEI Precipitation

At 0.15% PEI all the nucleic acids were precipitated from the sample; however, most of the Sac7d-IEP was also precipitated. Since the Sac7d-IEP with GTV linker eluted from the heparin column at a high NaCl concentration, it was decided that PEI precipitation would be carried out with an increased level of NaCl. The cell pellet from a 250ml expressed culture was re-suspended in heparin-bind buffer (0.7M NaCl) with protease inhibitor cocktail and loaded onto a heparin column. The protein was eluted from the column with a 20cv gradient of 0.7-2M NaCl. As before, the Sac7d-IEP was eluted at 1.4M NaCl. Samples from before PEI precipitation, the column load and flow-through along with the peak protein fractions were electrophoresed on a protein gel (Figure 4.31a and b).

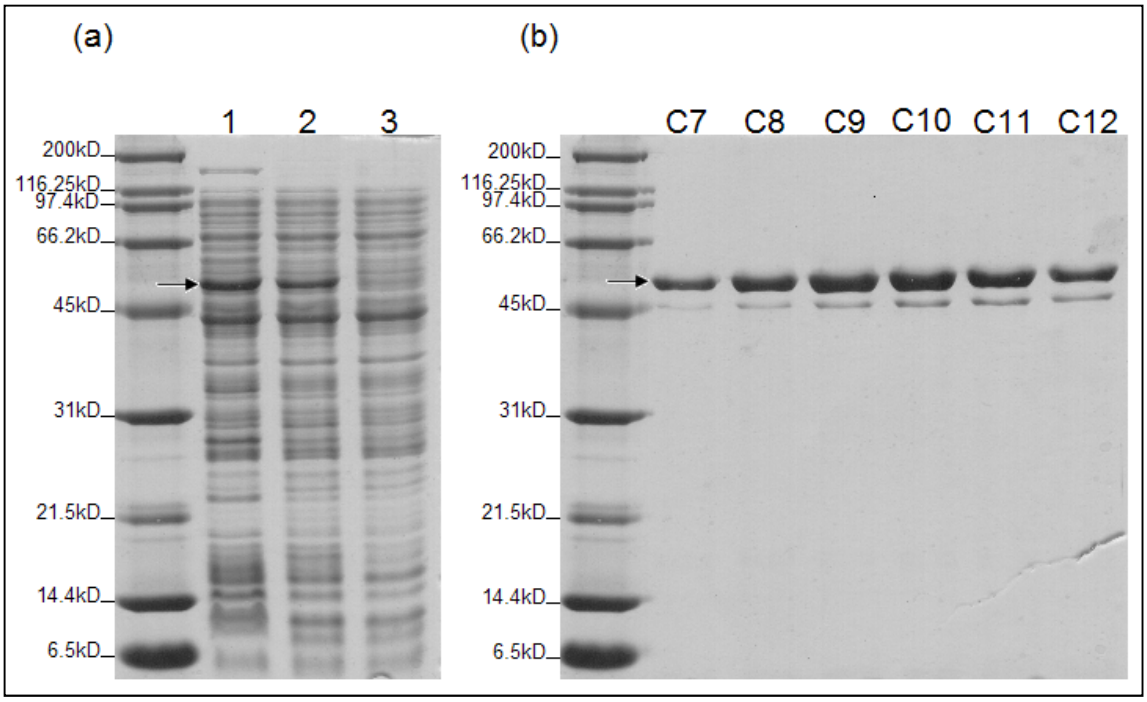

Figure 4.31:

a) A 12.5% SDS PAGE gel showing lane 1: the total protein sample before PEI precipitation, lane 2: PEI precipitated and filtered column load, and lane 3: the flow-through from the heparin column.

b) 12.5% SDS PAGE gel showing 15µl fractions of the protein peak containing Sac7d-IEP eluted at 1.4M NaCl.

Despite removing the linker from between the Sac7d and IEP domains, there was still an additional product co-purifying with the full-length protein.

Mass spectrometry analysis of this co-purify contaminant would reveal its identity and allow new approaches to be made to try to remove it or prevent it from occurring. If it was a random co-purifitying protein then efforts would need to be made to remove it, preventing its interference with the enzymatic activity of the Sac7d-IEP. However, It was expected that this product was due to either low level proteolytic cleavage, removing the Sac7d domain, and leaving a small percentage of wild type IEP or due to low level expression of the wild type IEP, without the Sac7d domain, from to a start codon further downstream. These were considered most likely due to the absence of this contaminant in the IEP-Sac7d purification. In either of these scenarios the Sac7d-IEP would be contaminated with wild type IEP which could show RT activity. Therefore, without removing this contaminant, it would not be possibly to separate activity and enzyme characteristics from the two proteins and no further work was carried out on this fusion product.

T. carboxydivorans IEP Purification

The IEP from *T. carboxydivorans* could not be expressed using TB medium instead LB medium was used for all large-scale expression with the OD_{600} reaching 6.7.

Heat-Treatment

In order to remove some background *E. coli* proteins, it was decided to heat treat the sample in the hope that the IEP from the thermophilic bacterium would remain soluble. Cells from 10ml of the over-expressed culture were harvested

and re-suspended in 1ml of cell lysis buffer D. The sample was sonicated and 100µl aliquots of the supernatant heated to 50℃ for 5, 10 and 15min and the soluble fraction loaded on a protein gel (Figure 4.32).

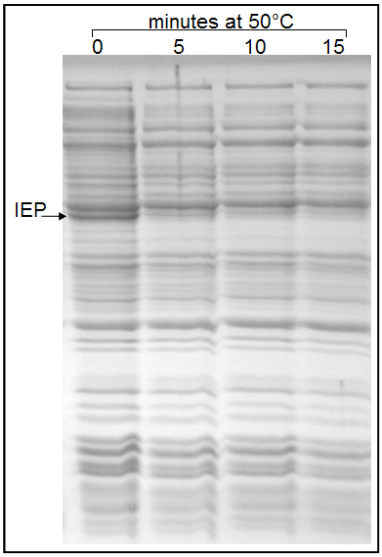

Figure 4.32: A 12.5% SDS PAGE gel showing the effect of 50℃ incubation on the soluble IEP.

Despite the fact that the *T.* carboxydivorans has an optimum growth temperature of 55℃, the IEP could not be heated at 50℃ for 5min without the protein becoming insoluble. Therefore, heat treatment could not be used to reduce the background level of *E. coli* proteins.

Nickel Column

A cell pellet harvested from 100ml of expressed culture was re-suspended in 5ml his-bind buffer A. The sample was sonicated, the soluble fraction filtered and loaded onto a 1ml bench-top cellulose column charged with Ni^{2+}. The flow-through and wash were collected and proteins eluted from the column using a step gradient of 5, 7, 14, 40 and 100% his-elute buffer B. Each step of the gradient was collected in 5x 1ml fractions. 5µl samples of the load, flow-

through, wash and third fraction from each step of the gradient were electrophoresed on a protein gel (Figure 4.33).

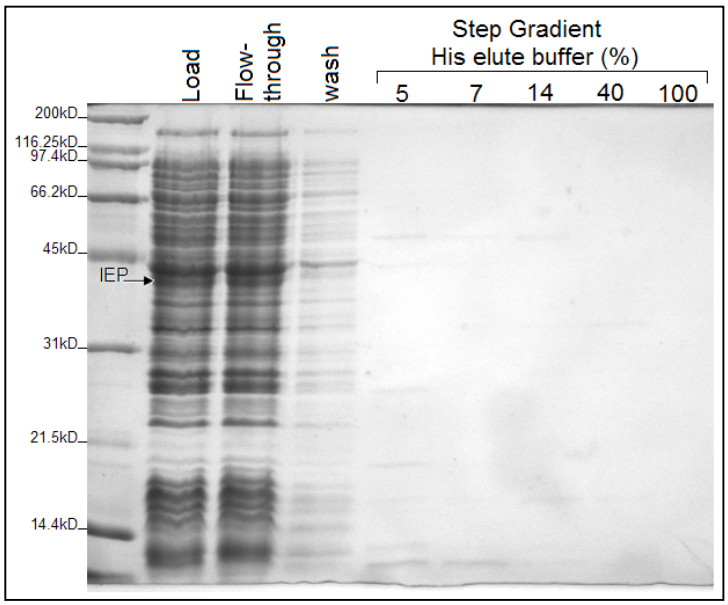

Figure 4.33: A 12.5% SDS PAGE gel showing the load and flow-through of the over-expressed *T. carboxydivorans* IEP loaded on a 1ml bench-top nickel column and samples of the elution fractions from the step gradient.

The protein gel of the load and flow-through showed that none of the IEP actually bound to the nickel column and came straight out in the collected flow-through.

PEI Precipitation

Nucleic acids could be interfering with the purification of the *T. carboxydivorans* IEP as is seen with the IEP from *B. caldovelox*. A cell pellet from large-scale protein expression was removed and re-suspended in heparin buffer A equivalent to 1ml/0.1g of cells. The sample was sonicated and 100µl aliquots treated to varying concentrations of PEI and incubated for 1h at 4ºC. After incubation, 10µl of the soluble fractions were electrophoresed on an agarose gel

to see the concentration required to remove all nucleic acids (Figure 4.34a), and onto a protein gel to see the effect of the PEI on the soluble IEP (Figure 4.34b).

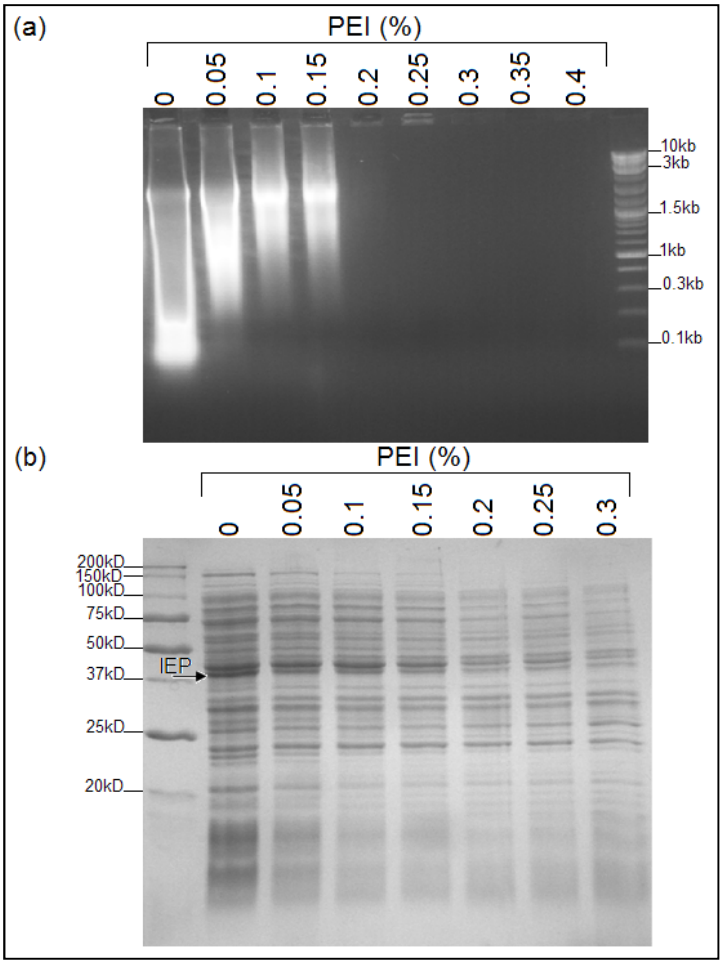

Figure 4.34:

a) A 1%(w/v) agarose gel showing the removal of nucleic acids from the expressed IEP culture with increasing concentrations of PEI

b) A 12.5% (v/v) SDS PAGE gel of the effects of increasing PEI on the IEP in the soluble fraction.

The agarose gel showed that 0.2% PEI was required to precipitate all the nucleic acids from the sample; however, at this concentration all the IEP had also precipitated. It was therefore necessary to repeat the experiment with 0.2% PEI with varied concentrations of NaCl. After incubation at 4ºC for 1h,

samples of the soluble fraction were electrophoresed on a protein gel (Figure 4.34).

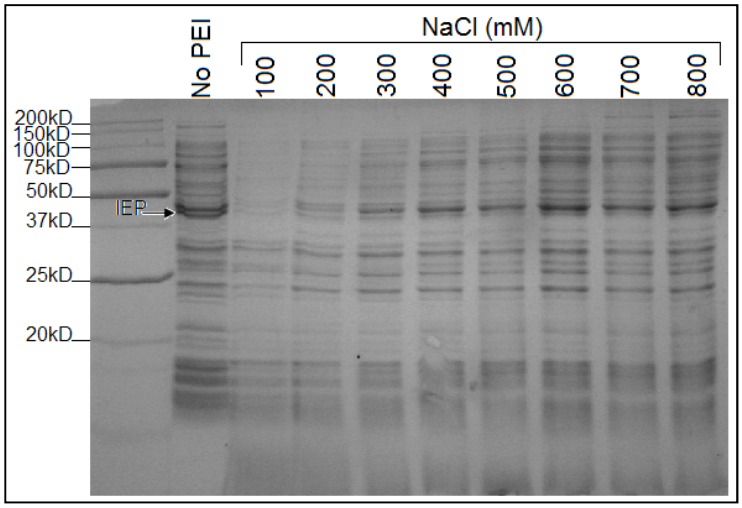

Figure 4.34: A 12.5%(v/v) SDS PAGE gel showing the effect of PEI on the soluble fraction of the IEP expression culture with increasing concentrations of NaCl.

The protein gel of the soluble fraction from the PEI precipitation showed that even increasing the NaCl up to 0.8M did not prevent the IEP from also precipitating.

Heparin Column – After PEI Precipitation

Three attempts were made to purify the IEP after 0.2% PEI precipitation in the hope that a small fraction of the protein had remained soluble. In each case a pellet equivalent to 250ml of expressed culture was re-suspended in heparin-bind buffer equivalent to 1ml/0.1g of cells. The attempts used the same protocol but altered the concentration of NaCl present in the bind buffer. These included 600mM NaCl heparin bind buffer, as shown to work for the *B. caldovelox* IEP, and 800mM NaCl. No protein was seen to bind to the column and it is possible that, if any IEP had remained soluble, the protein could not bind the column at such high NaCl concentrations and would therefore come straight out in the

flow-through. Additionally, an attempt was made with 100mM NaCl. Although it was expected that all the protein would be precipitated with the PEI exposure at such a low NaCl concentration it was thought that, if any protein had remained soluble, this NaCl concentration should be low enough to allow the IEP to bind and be eluted from the heparin column. Unfortunately all three of these attempts at purifying the IEP using the heparin column failed.

Q Column

In theory, the IEP from *T. carboxydivorans*, with a pI value of 10.35, should not bind to a Q column. However, nucleic acids interfered with the charge of the *B. caldovelox* IEP allowing it to bind to this anion exchange column. It was therefore decided to load the *T. carboxydivorans* IEP onto a Q column in the hope that it would be eluted in separate fractions from the major nucleic acid peaks. The cell pellet from a 250ml expressed culture was re-suspended in Q-bind buffer, sonicated and the soluble fraction filtered. This was then loaded onto a 5ml Q column and the flow-through collected. Proteins were eluted from this column using a 0.025-2M NaCl gradient over 20cv with 1.4ml fractions collected. Samples of the fractions for each protein peak observed on the trace (Figure 4.35) were tested for RT activity.

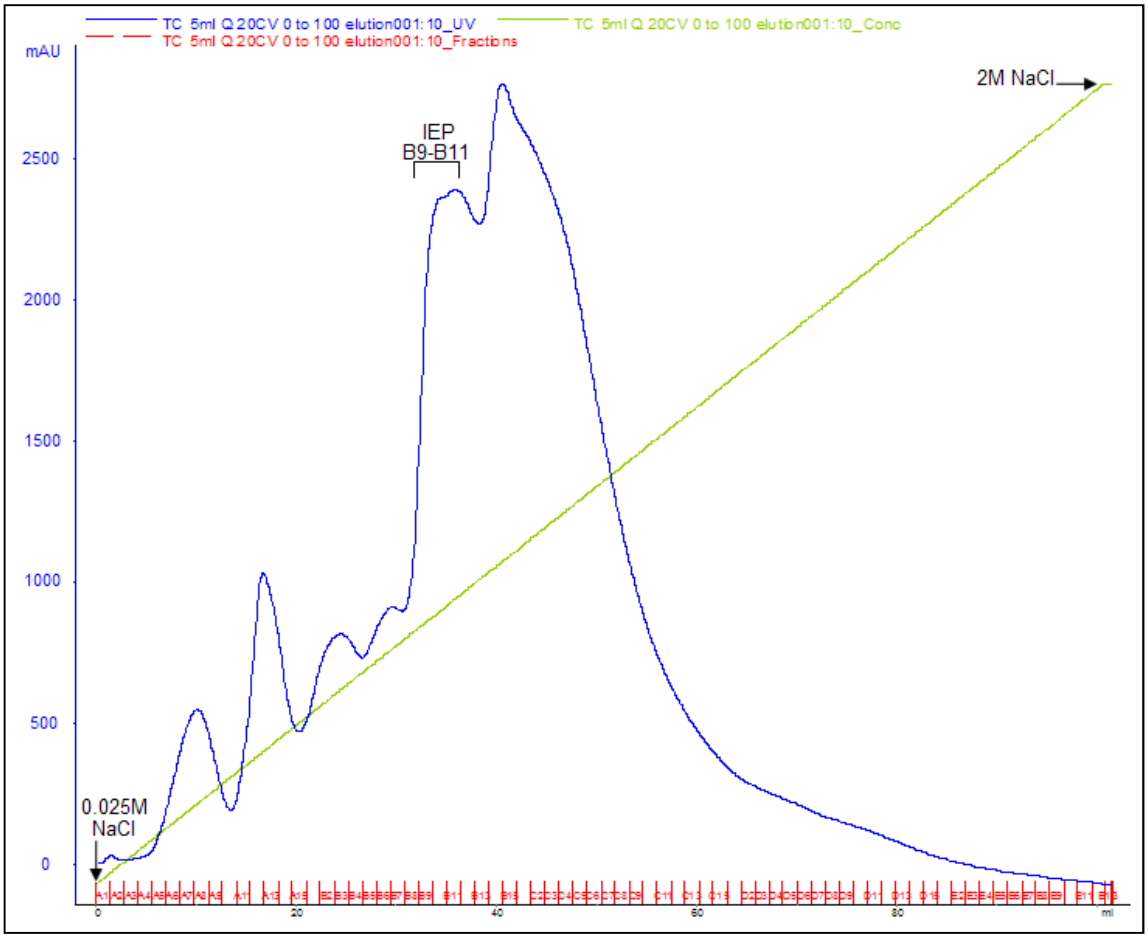

Figure 4.35: Purification trace of the IEP expression sample loaded onto a 5ml Q column. Proteins were eluted using a gradient of 0.025-2M NaCl.

0.25µl of each peak fraction was used as the enzyme in a cDNA synthesis step using 20ng MS2 RNA and MS2:3395_R (Appendix I) and incubation for 30min at 45ºC. An aliquot of this reaction was then used as a template in a PCR with MS2:3395_F and MS2:3231_R. 5µl of this reaction was electrophoresed on an agarose gel (Figure 4.36).

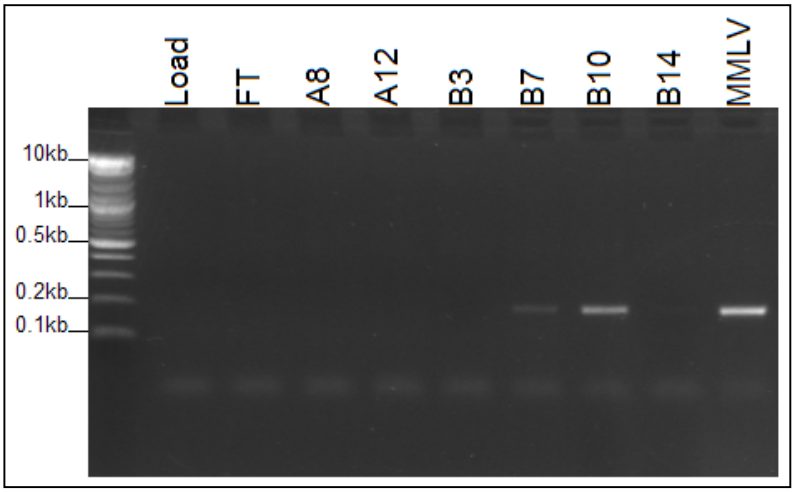

Figure 4.36: A 2%(w/v) agarose gel showing the amplification of cDNA generated using aliquots of the peak protein fractions as seen on the trace.

The agarose gel showed that there was no activity in the flow-through and that the most activity was seen in the A_{280} peak B10 which eluted at 665mM NaCl. Fractions across this peak were electrophoresed on an SDS PAGE gel (Figure 4.37a) to analyse the purity of the protein and onto an agarose gel to assess the level of nucleic acids present in each fraction (Figure 4.37b).

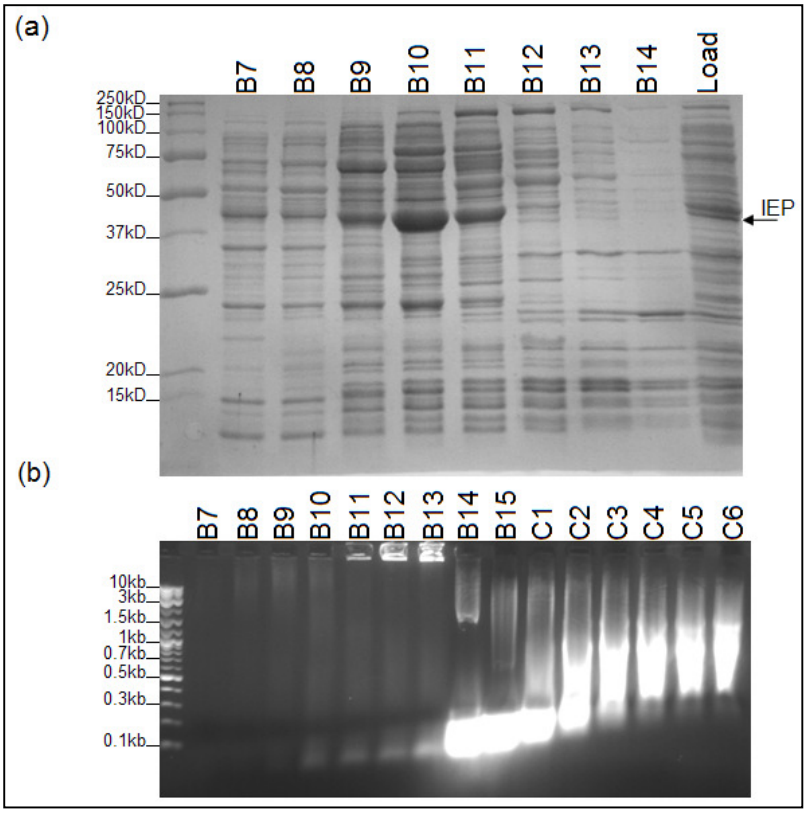

Figure 4.37:

a) 12.5% (v/v) SDS PAGE gel showing 5µl aliquots of the fractions from the protein peak with RT activity.

b) A 1%(w/v) agarose gel showing nucleic acids within the main RT activity peak.

The protein gel showed that the IEP eluted with a high proportion of additional *E. coli* proteins. However, the agarose gel showed that the main three fractions, B9-11, eluted with a low percentage of contaminating nucleic acids. It was hoped that enough nucleic acids had been removed from the sample to allow additional purification steps.

Heparin Column – After Q Column

Fractions B9-B11 were pooled and dialysed into heparin-bind buffer (25mM NaCl). Once filtered, the sample was loaded onto a 5ml heparin column, the

flow-through collected and proteins eluted using a 0.025-2M NaCl gradient over 20cv. Protein peaks were present at the beginning of the trace, between 140-660mM NaCl, but no peak was seen at higher NaCl concentrations where other IEPs had previously eluted (Figure 4.38).

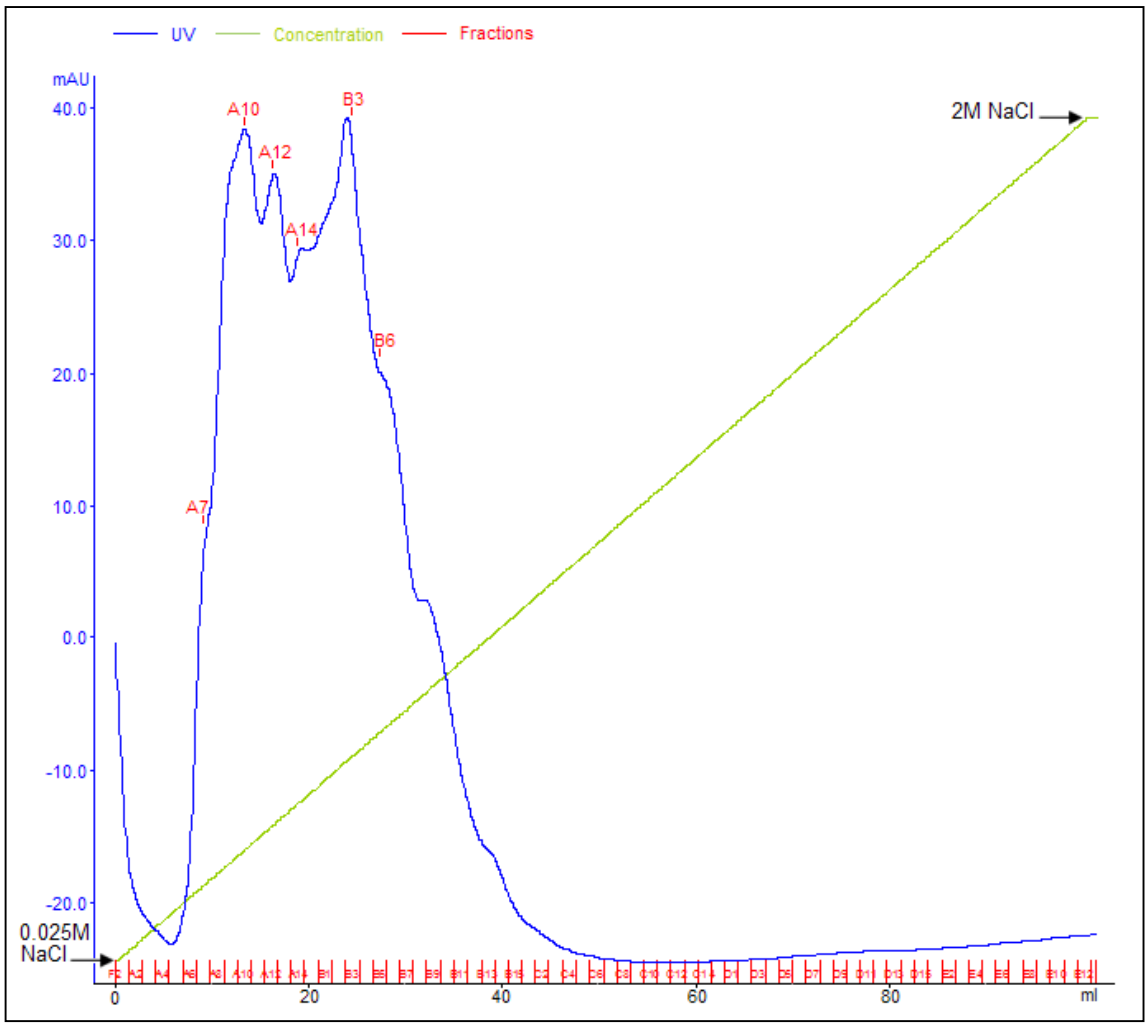

Figure 4.38: A purification trace showing the elution of *T.* carboxydivorans IEP from a 5ml Heparin column after partial purification using a Q column. Proteins were eluted using a gradient of 0.025-2M NaCl. Main protein peaks, later electrophoresed on a protein gel, are labelled in red.

Fractions from each protein peak were electrophoresed on a protein gel to assess where the IEP eluted (Figure 4.38a) and each peak was assayed for RT

activity by using 0.25µl of each fraction as the enzyme in a DNA synthesis step followed by a PCR to amplify the cDNA (Figure 4.38b).

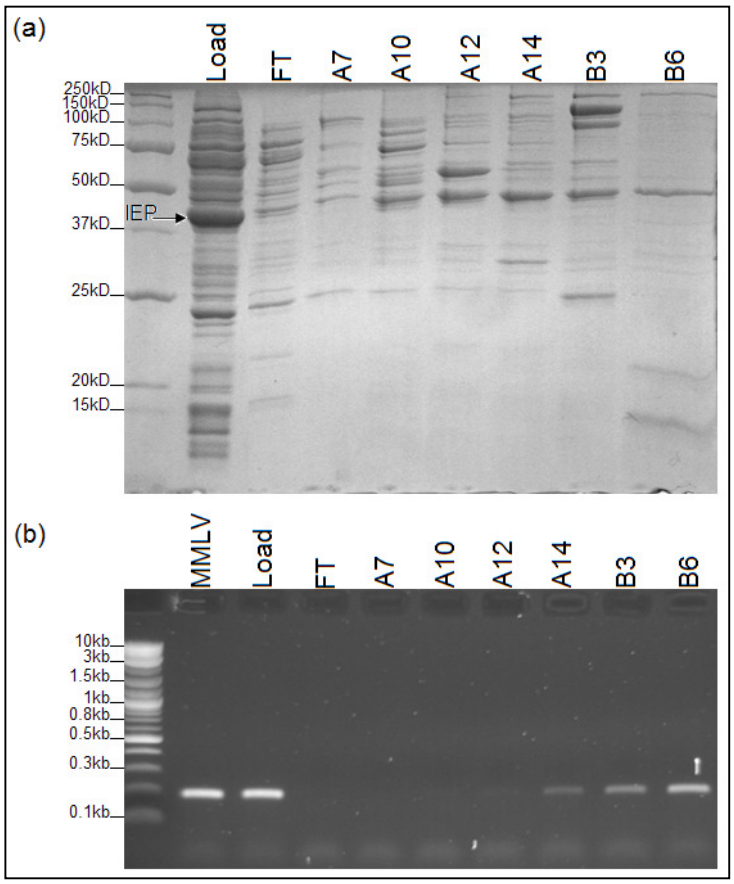

Figure 4.38:

a) A 12.5%(v/v) SDS PAGE gel showing 15µl samples of the peak fractions as seen on the purification trace.

b) A 2%(w/v) agarose gel showing RT activity of the peak protein fractions.

The protein gel showed that most of the peaks contained some of the IEP as it had eluted very broadly with other proteins. However, the main activity was seen in the later peaks in A14-B6. These fractions did not appear to contain more IEP than the other fractions and it was possible that there was something inhibitory in the earlier fractions or that the enzyme was more active in higher concentrations of NaCl.

SP Column – After Q Column and Heparin Column Purification

Fractions A14-B6 were pooled and dialysed into SP-bind buffer and loaded onto a 1ml SP column. Proteins were eluted from the column using a gradient of 0.025-2M NaCl over 100cv and 1ml fractions were collected. All the proteins eluted from the column at the beginning of the gradient between 100-525mM NaCl (Figure 4.39).

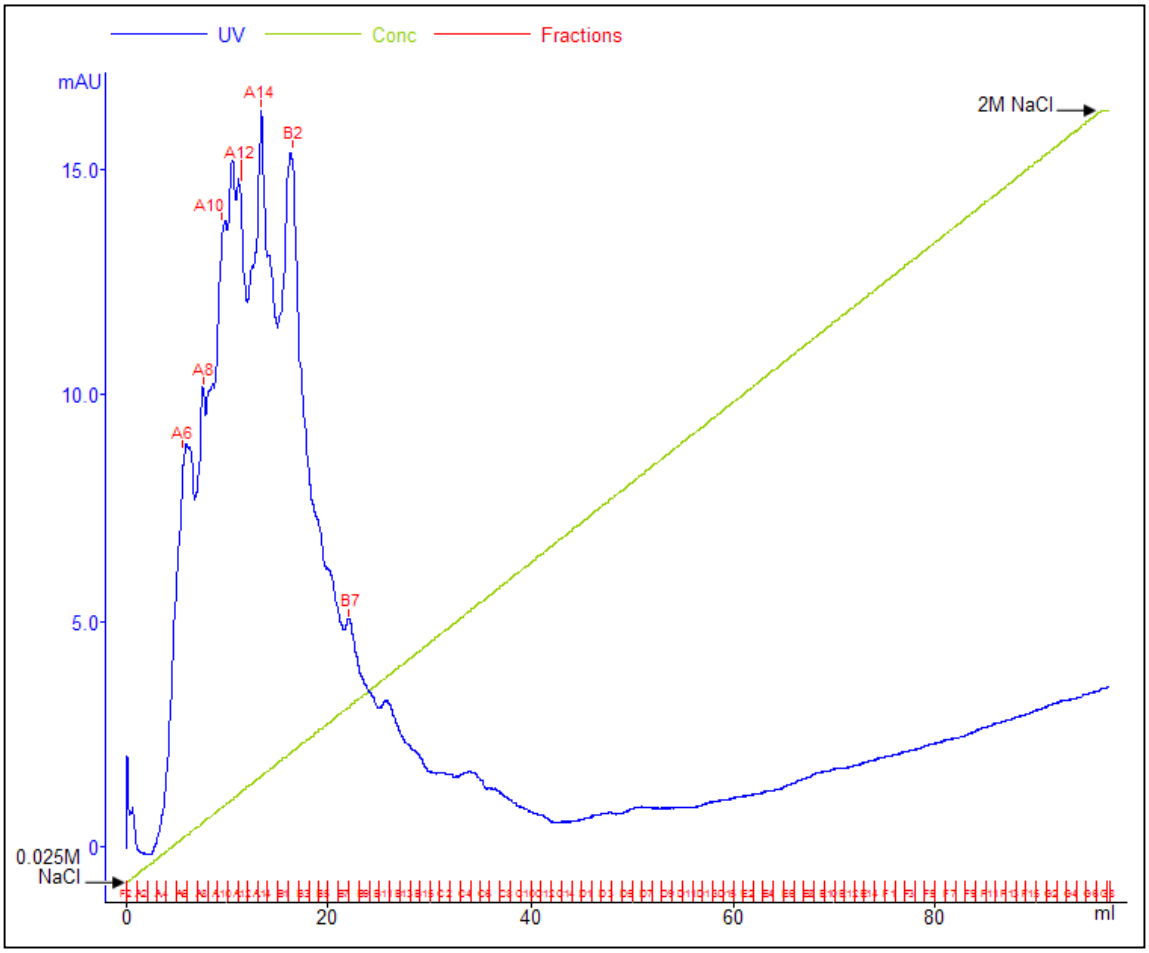

Figure 4.39: Purification trace showing the elution profile of IEP on an SP column after Q and heparin purification. Main protein peaks, later electrophoresed on a protein gel, are labelled in red.

15µl of each of the main protein peaks were electrophoresed on a protein gel which showed that the IEP eluted very broadly and was found across all of the peaks (Figure 4.40).

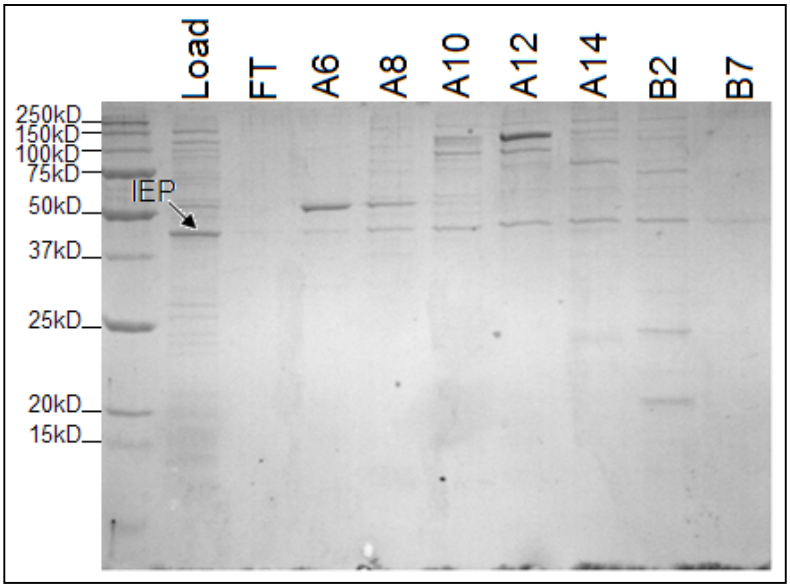

Figure 4.40: A 12.5%(v/v) SDS PAGE gel showing 15µl samples of the main protein peaks of the elution of the IEP from the SP column.

The three purification steps performed on the *T. carboxydivorans* IEP did not result in pure protein and each step reduced the overall yield. Therefore alternate purification methods were required.

<u>Ni^{2+} Charged Sepharose Column – After Q Column Purification</u>

A second attempt was made to utilise the N-terminal his-tag on this IEP. It was hoped that the Q column would separate the IEP from the majority of the nucleic acids, thereby preventing their interference with the protein binding to the Ni^{2+} charged column. The IEP was purified using a Q column as before and the fractions containing the IEP were pooled and dialysed into his-bind buffer C. A 14ml IDA Sepharose column was charged with Ni^{2+} and equilibrated with his-bind buffer C. The partially purified IEP was loaded onto the nickel column and

the protein eluted with a gradient of 5-300mM Imidazole over 10v. One peak was seen on the purification trace (Figure 4.41) corresponding to an elution of proteins at 100mM imidazole.

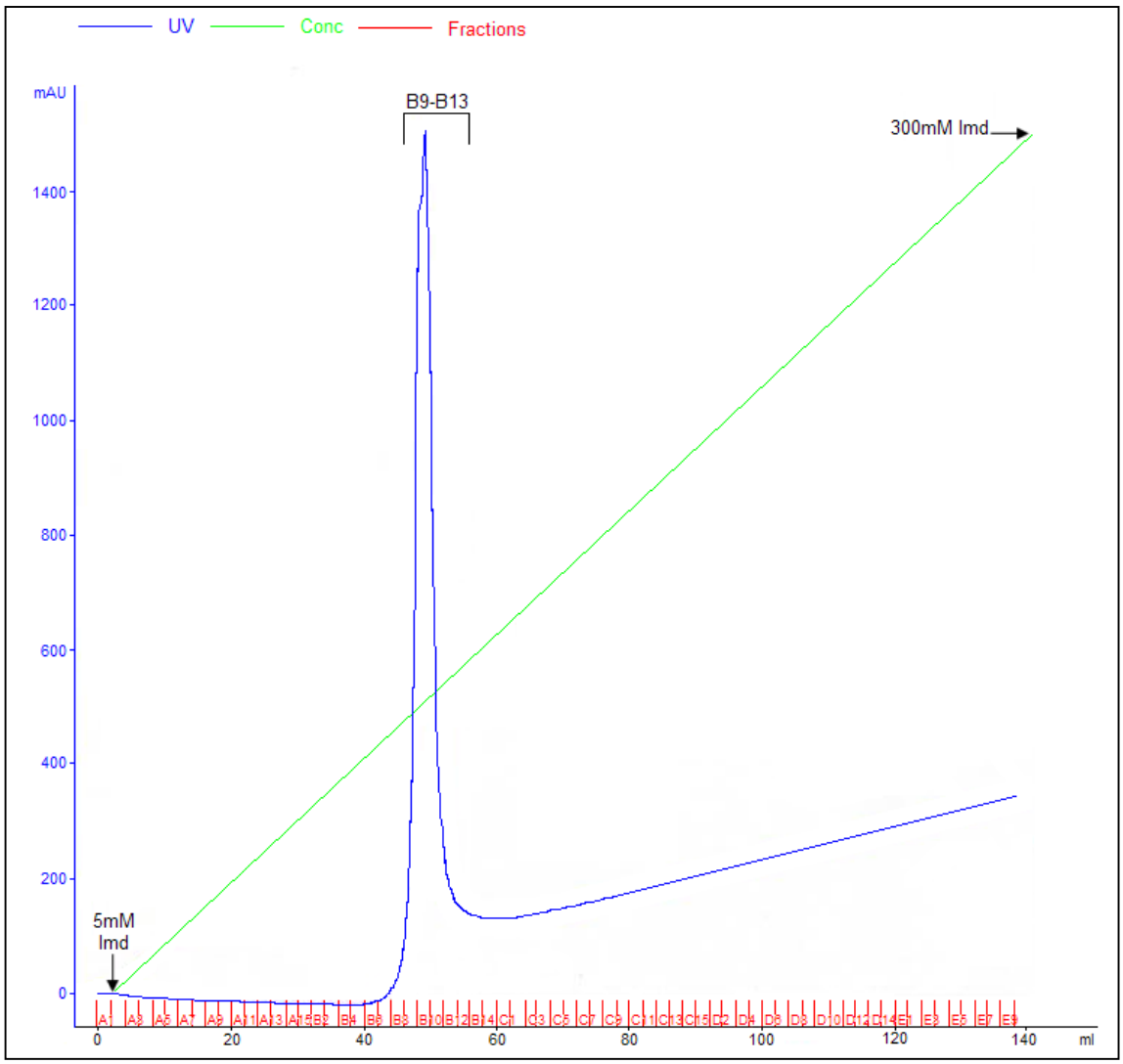

Figure 4.41: A purification trace showing the elution profile of *T. carboxydivorans* IEP from a 15ml Ni^{2+} charged IDA Sepharose column after partial purification using a Q column. The protein was eluted using an imidazole (Imd) gradient of 5-300mM.

Fractions from the protein peak were electrophoresed on a protein gel to assess the purity of the sample (Figure 4.42a) and 0.25µl used in a reaction to assay

the RT activity (Figure 4.42b). The protein gel showed that the IEP eluted within this peak but was also contaminated with the original proteins loaded on to the column. These fractions did show RT activity with the peak activity being found in the peak fraction B10. However, the sample still contained contaminating proteins that could interfere with the assay of enzyme activity and therefore prevent accurate characterisation of the enzyme.

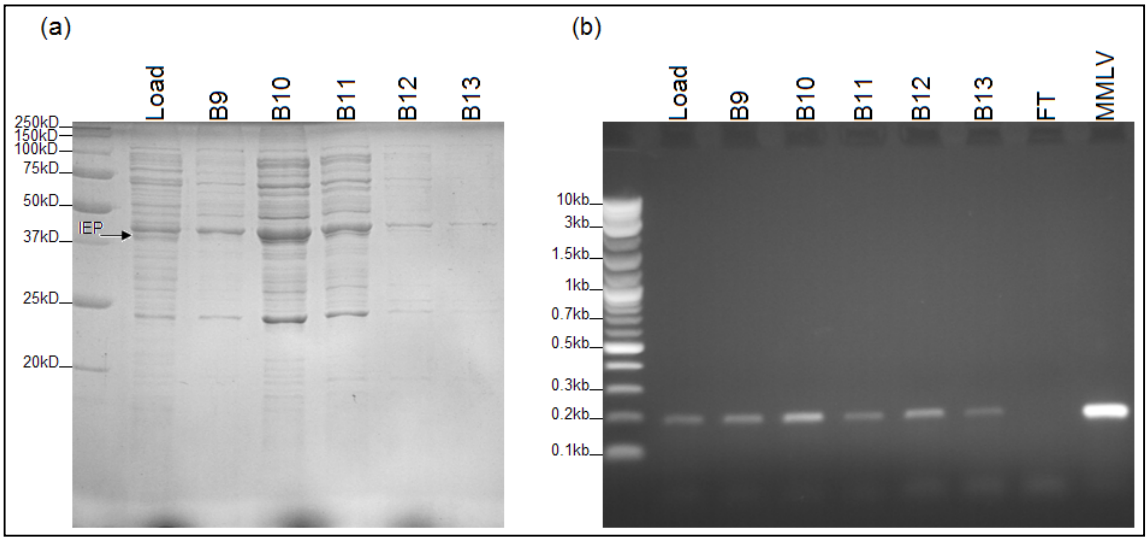

Figure 4.42:
a) 12.5%(v/v) SDS PAGE gel showing the fractions from the elution peak of the nickel charged column.
b) RT activity in the fractions from the protein peak.

SP Column – After Q Column

It was hoped that by removing some of the nucleic acids from the protein sample, by use of a Q column, that the IEP would then bind to an SP column as would be expected with the high pI value of the protein. A cell pellet harvested from a 250ml expressed culture was re-suspended in 5ml Q-bind buffer, sonicated and filtered. The IEP was loaded and eluted from the Q column, as before, and fractions containing the IEP were pooled and dialysed into SP-bind buffer, filtered and loaded onto 1x 5ml SP columns. Samples of the pre-filtered

load, the column load and the flow-through were analysed on a protein gel which showed the IEP to be present in the flow-through and not to bind to the column (Figure 4.43). It was likely that the nucleic acids that were still present were affecting its ability to bind to the SP column.

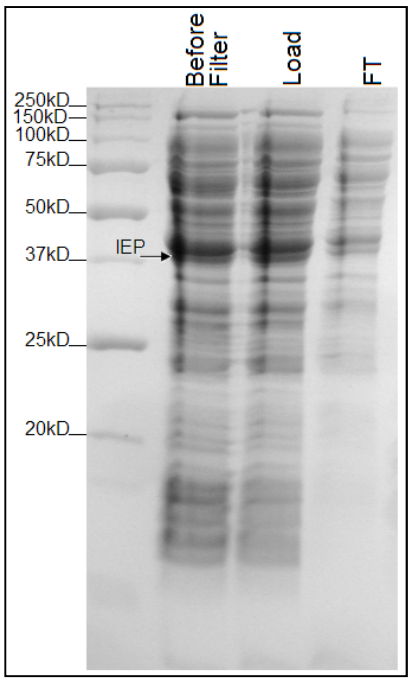

Figure 4.43: A 12.5% SDS PAGE gel showing the load and flow-through collected from an SP column after loading with an IEP expression culture previously partially purified on a Q column.

Ceramic Hydroxyapatite Column

A cell pellet equivalent to 250ml cell culture was re-suspended in 20ml CHT-bind buffer, sonicated, filtered and loaded onto a 19ml CHT column. The flow-through was collected and proteins eluted from the column using a gradient of 20-400mM potassium phosphate pH6.8 over 10cv. The purification profile showed mainly one broad protein peak (Figure 4.44).

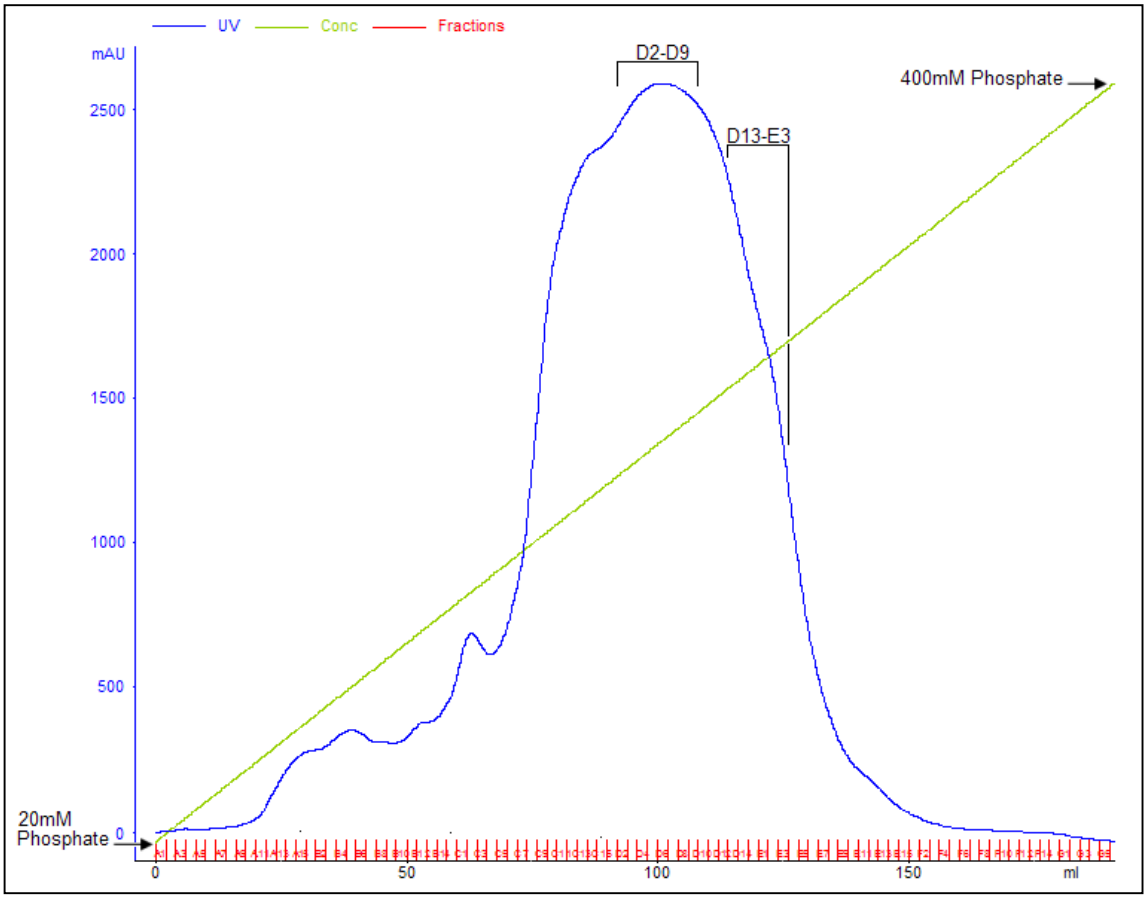

Figure 4.44: A purification trace showing the elution profile of the IEP protein expression sample from a CHT column with an elution gradient of 20-400mM potassium phosphate, pH6.8.

Initially, 0.25µl samples of fractions within the protein peaks were assayed for RT activity and two peaks of activity were found (Figure 4.45).

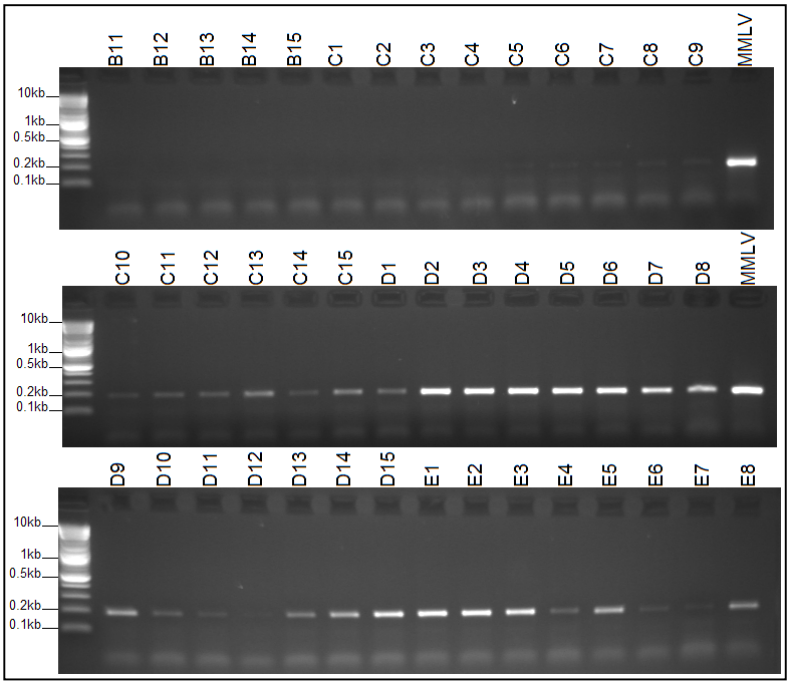

Figure 4.45: A 1%(w/v) agarose detecting the amplification of cDNA produced using 0.25µl of the fractions as the RT enzyme, 20ng MS2 RNA and MS2:3395_R and with MS2:3231_F in the PCR.

Fractions from the activity peaks were electrophoresed on a protein gel to analyse the IEP purity (Figure 4.46).

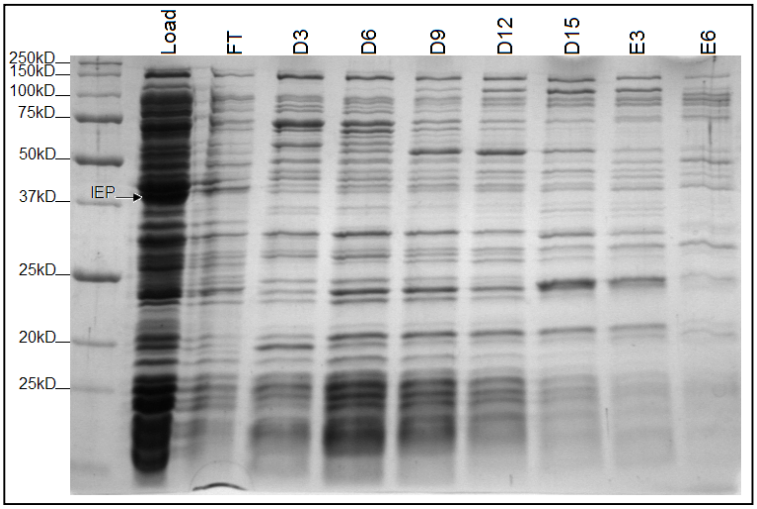

Figure 4.46: A 12.5%(v/v) SDS PAGE gel showing 5µl samples of the fractions with RT activity.

None of the fractions that showed RT activity had a large amount of the expected size IEP. By running protein samples out on an agarose gel (Figure 4.47) it was revealed that the activity seen actually corresponds to where nucleic acids were eluted from the column. It was hypothesised that the activity seen was in fact due to a low level of IEP that associates and elutes with the nucleic acids.

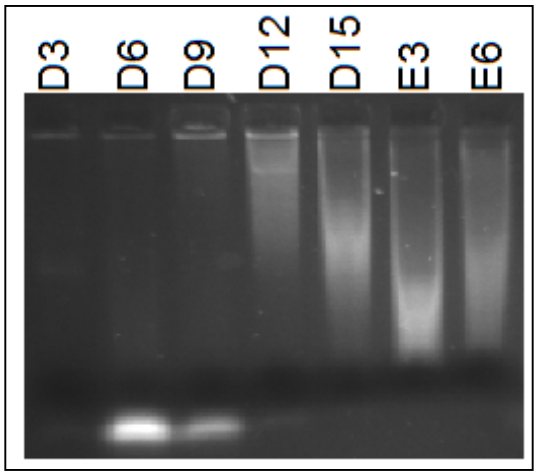

Figure 4.47: A 1%(w/v) agarose gel showing the elution of nucleic acids from the CHT column.

In order to locate the main protein peak it was necessary to load fractions on a protein gel instead of using the activity assay (Figure 4.48).

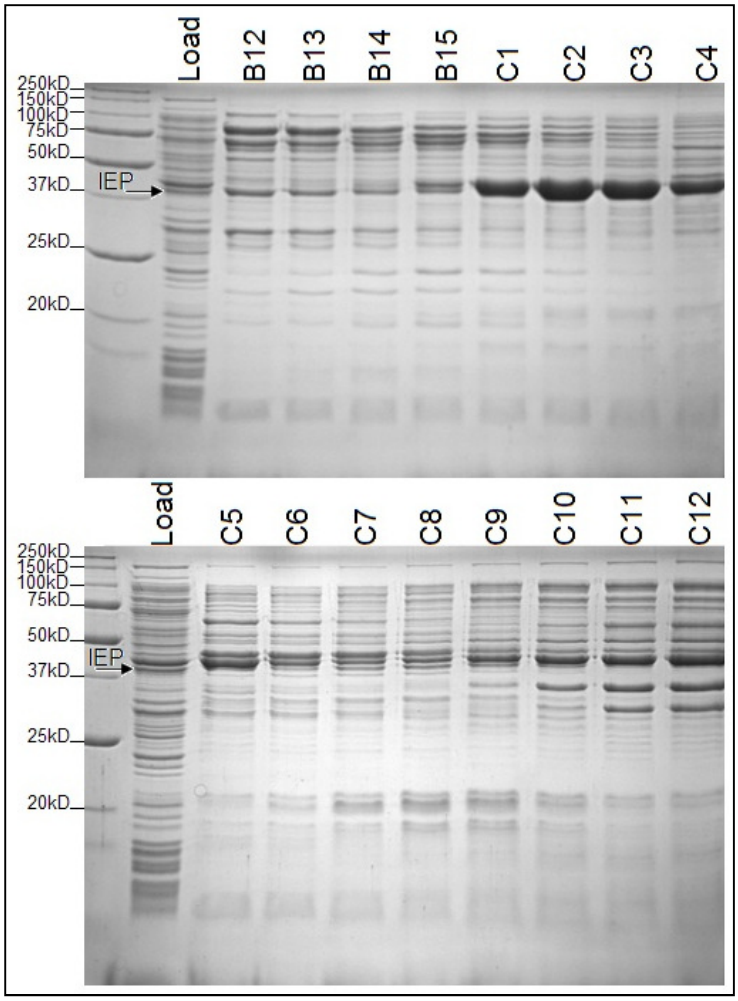

Figure 4.48: A 12.5%(v/v) SDS PAGE gel showing 5µl aliquots of the fractions from the elution of the IEP from the CHT column.

By analysis of the protein gel alone it was difficult to assess where the protein had eluted and there was a possibility that it had eluted at either 120mM phosphate at fractions B12,-13 140mM phosphate at fraction C1-5 or 180mM phosphate at fraction C10-12. A second RT assay reaction was performed, diluting the fractions further to enable inhibitors to be reduced and 40 PCR cycles carried out (Figure 4.49).

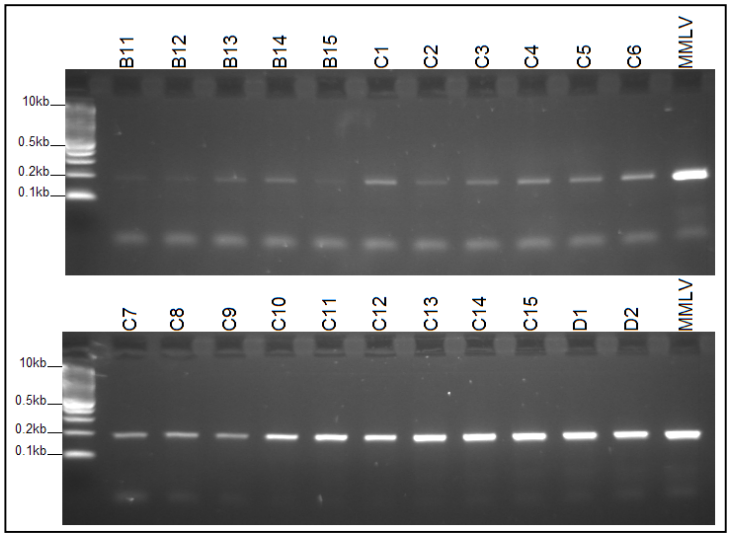

Figure 4.49: A 2%(w/v) agarose gel showing the result of an RT assay on fractions believed to contain the IEP. The assay shows the amplification of cDNA synthesised from the MS2 RNA template.

The results of the assay revealed very low level of activity in the expected fractions B11-14 and higher activity in the later fractions. It is possible that there are still inhibitors present in these earlier fraction affecting their activity. The fact that there are clear activities across the fractions, even where no clear IEP is present, suggest that this enzyme is in fact very active, detectable by the sensitivity of the assay. Unfortunately the fractions where the IEP was detected were not stable, even at 4ºC, and after several hours a large volume of precipitate had formed and these fractions no longer showed any RT activity.

CHT Purification – After PEI Precipitation

Since CHT columns are not affected by high NaCl concentration, it was decided to attempt to precipitate the nucleic acids at 2M NaCl before CHT column purification. A cell pellet from a 250ml expressed culture was re-suspended in CHT-bind buffer with 2M NaCl at the equivalent of 1ml/0.1g of cells. PEI at 0.2%(w/v) was used to precipitate the nucleic acids. It was not possible to assess whether the precipitation of the nucleic acids had been successful, as

adding the sample to an agarose gel loading buffer reduced the NaCl concentrations resulting in protein precipitation within the sample. After centrifugation and filtering the PEI precipitated sample was loaded onto a CHT column and proteins eluted using an phosphate gradient of 20-400mM over 10cv. The purification trace (Figure 4.50) revealed three protein peaks and samples from each peak were assayed for activity.

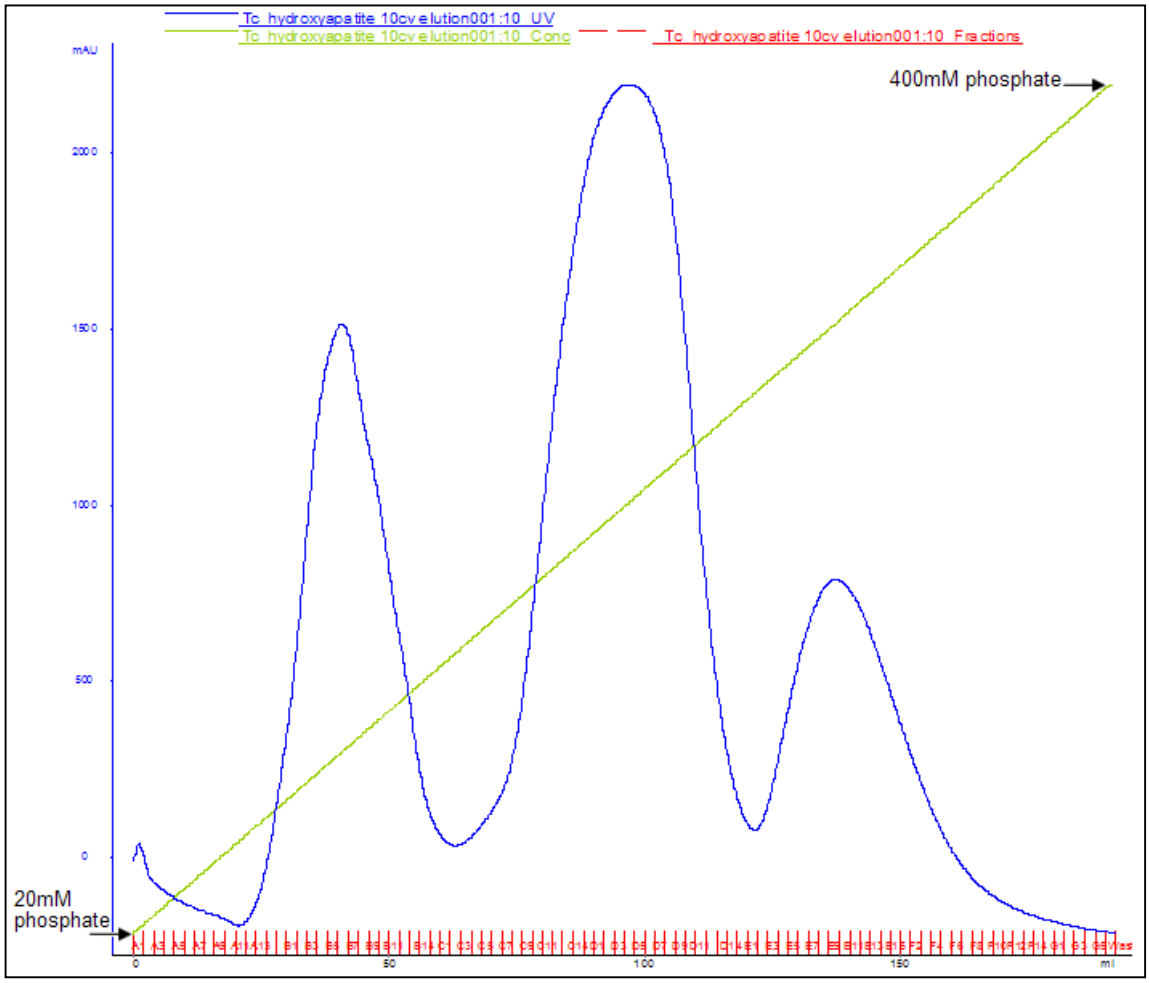

Figure 4.50: The purification trace showing the elution profile of an IEP expression sample after precipitation with PEI. Proteins were eluted from the column using a gradient of 20-400mM potassium phosphate.

There were only three protein peaks present on the purification trace. The first major peak corresponded to fractions A14-C1, peaking at B7, the second spans

C9-D15, peaking at C3, and the third spanning E3-F6 peaking at E9. The main activity was seen in the protein peak eluting at 80mM potassium phosphate with the major activity in B7 (Figure 4.51). However, the activity fluctuated across the fractions and was seen in all the protein peaks

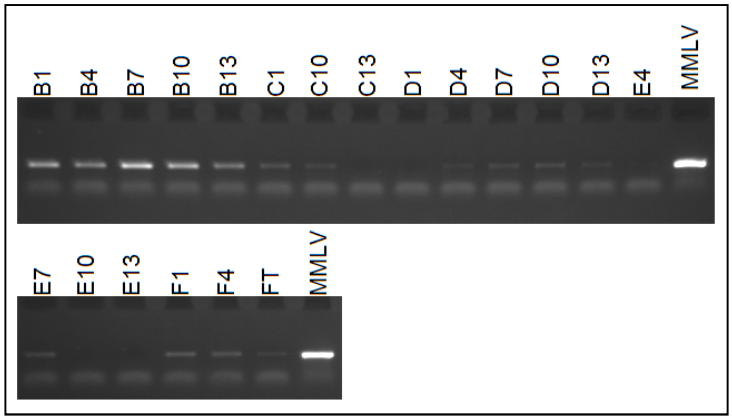

Figure 4.51: A 2%(w/v) agarose gel showing the level of RT activity within each fraction when 0.25µl was used to synthesise cDNA from a MS2 RNA template.

When the a sample of the column load was run on the gel (Figure 4.52) it showed a very low level of IEP possibly due to the fact that the IEP had precipitated during the PEI step. The fractions corresponding to the major activity, B7-14, were also electrophoresed on the same protein gel there was the presence of a partially purified protein but with comparison to the IEP present in the load, it appeared to be slightly larger than expected (Figure 4.52).

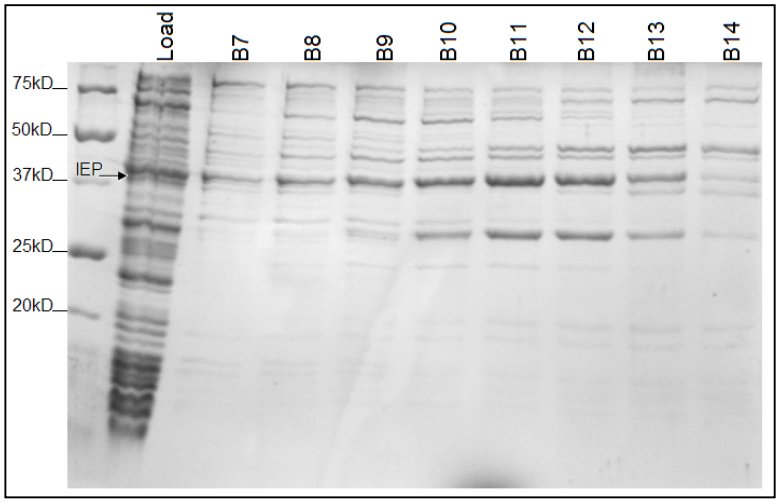

Figure 4.52: 12.5% SDS gel with samples of the fractions showing the peak RT activity.

The fractions that did show activity were loaded on an agarose gel and revealed a high degree of nucleic acid contamination, suggesting that the PEI step had not been completely successful (Figure 4.53). It is likely that the activity seen is due to a low level of IEP that becomes associated and elutes with the nucleic acids.

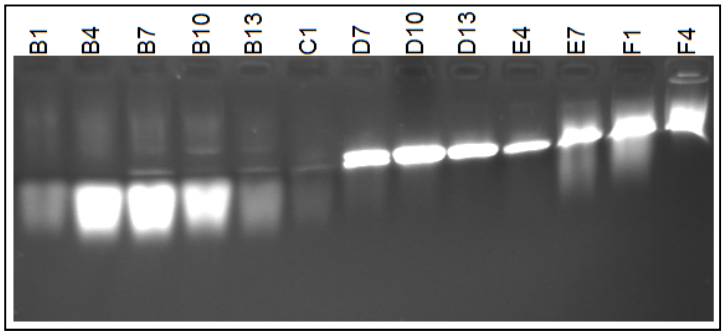

Figure 4.53: A 1%(w/v) agarose gel showing the nucleic acid contamination of fractions from the elution of the IEP from the CHT column. Fractions B1-13 were previously shown to contain the peak activity.

In order to assess whether the IEP was associating with DNA or RNA or both, 8µl of the fraction was mixed with 1µl of the appropriate 10x buffer and 1µl of

either DNase or RNase and incubated at room temperature for 1h. The sample was then loaded on an agarose gel (Figure 4.54) which showed that the IEP activity seen elutes with the RNA fractions.

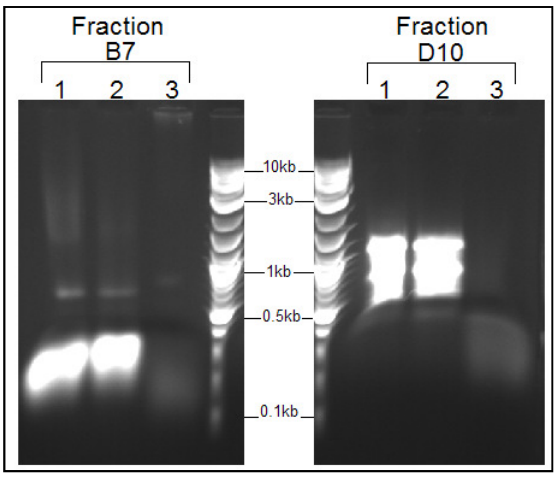

Figure 4.54: A 1%(w/v) agarose gel showing the effect on the nucleic acids in the fraction (lane 1) when exposed to DNase (lane 2) and RNase (lane 3).

No further attempts to purify the *T. carboxydivorans* IEP were carried out, and therefore no characterisation of the enzyme could be performed. However, the fact that the IEP did show activity showed that it has the potential, once purified, to be useful thermostable RT enzyme.

P. mobilis IEP2 Purification

Initially the *P. mobilis* IEP was over-expressed on a large-scale using TB medium. However, the protein gels from this expression culture revealed no protein at the expected size in either the soluble or insoluble fraction. Large-scale expression therefore had to be carried out using LB media where OD_{600} reached 2.6 and the expressed IEP was 100% soluble.

Heat Treatment

A sample of the large-scale IEP expression culture was removed, sonicated and centrifuged. The supernatant was then exposed to 50ºC for 5, 10 and 15min and the soluble fraction electrophoresed on a protein gel (Figure 4.55). The results of this gel showed that the IEP behaved the same as the previous IEPs and was not stable even after 5min at 50ºC. Heat treatment was therefore not an option for purification to remove some background *E. coli* proteins from the protein sample.

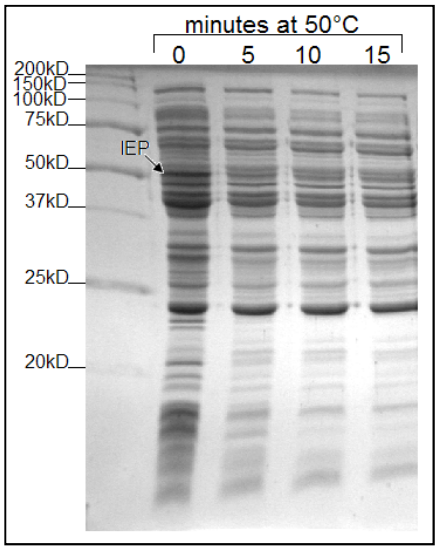

Figure 4.55: A 12.5%(v/v) SDS PAGE gel showing the effect of 50ºC heat treatment on the soluble fraction of the IEP.

Heparin Column

The cell pellet from a 250ml expression culture was re-suspended in 10ml heparin-bind buffer and loaded onto a 5ml heparin column. Protein was eluted from the column with a gradient of 0.025M-2M NaCl over 20cv. Samples of the load and flow-through were electrophoresed on a protein gel (Figure 4.56). This gel showed that the IEP did not bind to the column and was found in the flow-through. Presumably the IEP has a higher affinity for nucleic acids than for the

column and therefore could not bind. In order to use a heparin column it would be necessary to remove the nucleic acids from the sample.

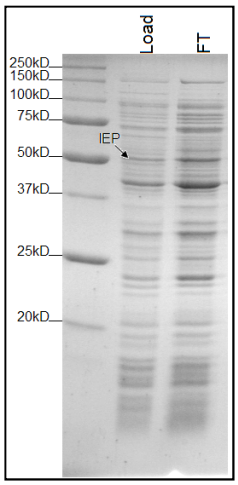

Figure 4.56: A 12.5%(v/v) SDS PAGE gel showing the IEP sample loaded onto a heparin column followed by the flow-through. The IEP did not bind the column and came out in the flow-through.

PEI Precipitation

A sample from the over-expressed IEP culture was removed and the pellet re-suspended in heparin-bind buffer equivalent to 1ml/0.1g of cells. The sample was sonicated and 100µl aliquots exposed to varying concentrations of PEI. After incubation, the soluble fractions were electrophoresed on an agarose gel to assess the level required to remove all nucleic acids (Figure 4.57a), and onto a protein gel to see the effect of PEI on the soluble IEP (Figure 4.57b).

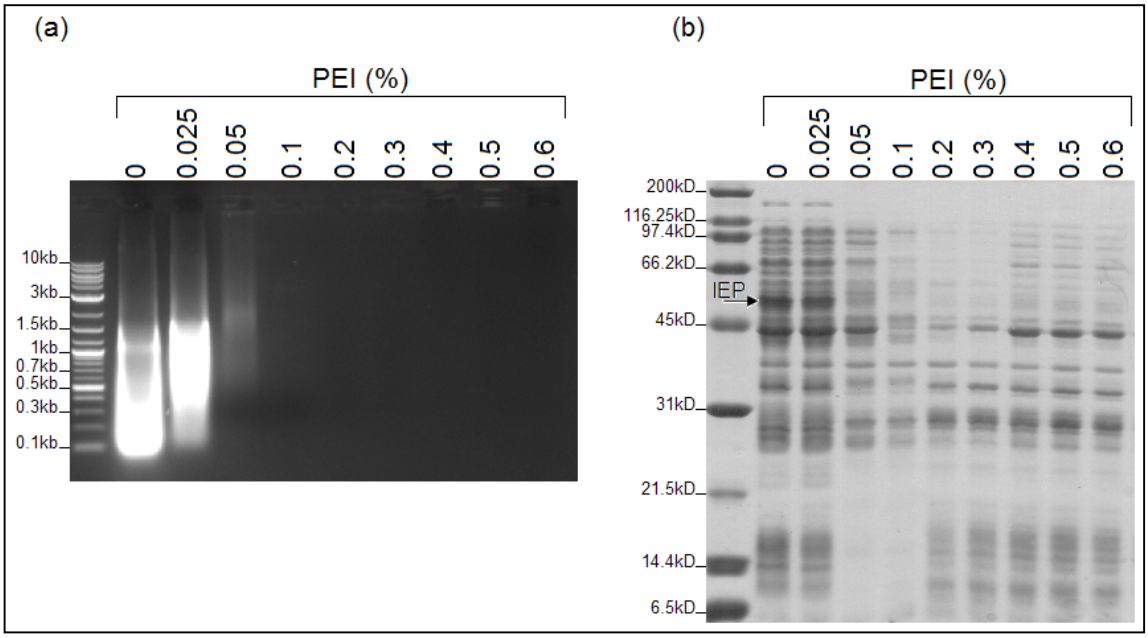

Figure 4.57:

a) A 1%(w/v) agarose gel showing the required level of PEI needed to remove all the nucleic acids from the IEP expression sample

b) A 12.5%(v/v) SDS PAGE gel showing the effect of increasing PEI concentrations on the soluble IEP.

0.1% PEI was required to remove all the nucleic acids from the sample; however, at this concentration, all the IEP was also precipitated. The experiment was repeated with a 0.1% PEI and varying the concentration of NaCl. After incubation, the soluble fraction was electrophoresed on a protein gel (Figure 4.58).

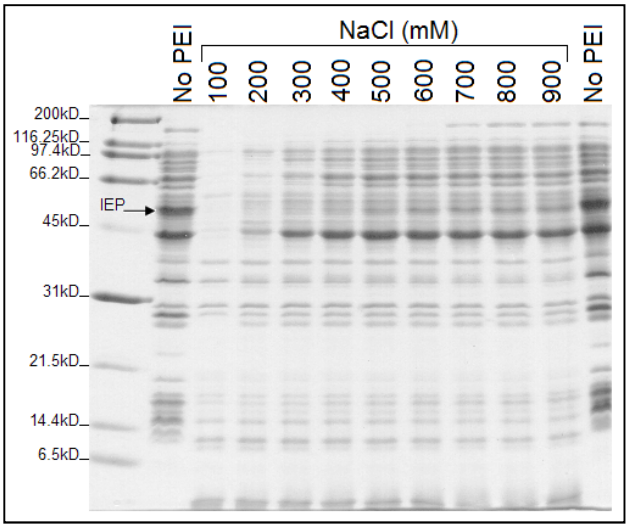

Figure 4.58: A 12.5%(w/v) SDS PAGE gel showing the effect on the IEP with 0.1% PEI and increasing concentrations of NaCl.

The protein showed that, even after 900mM NaCl was used, very little of the IEP remained soluble. Despite the very low level of soluble IEP it was decided to use this method to try to remove nucleic acids in the hope that any remaining IEP could still bind to a column and would allow enough IEP to be purified for characterisation.

Heparin Column – After PEI Precipitation

Initially it was decided to purify the IEP using the same method that was shown to work for the *B. caldovelox* IEP. A cell pellet from a 250ml expressed culture was re-suspended in heparin-bind buffer (600mM NaCl) equivalent to 1ml/0.1g of cells. The sample was sonicated and incubated with 0.1% PEI at 4ºC for 1h. The sample was centrifuged, the supernatant filtered, loaded onto a heparin column and eluted with a gradient of 0.6-2M NaCl over 20cv. Unfortunately, when electrophoresed on a protein gel, no IEP could be seen in either the load or any of the protein peaks from the purification trace. It was likely that all the IEP had been removed at the PEI precipitation step and the purification was repeated as before but with a heparin-bind buffer containing 900mM NaCl.

Again, no IEP could be seen in the load, flow-through or protein peaks. It is possible that if low levels of IEP remained soluble after the PEI precipitation that 900mM NaCl was too high to allow the IEP to bind and the protein would be present in the flow-through.

The heparin purification was also repeated with PEI precipitation carried out at 100mM NaCl. It was hoped that if any IEP remained soluble, 100mM NaCl would be low enough to allow the protein to bind to the column. However, there was no IEP present in the load, flow-through or any of the protein peaks.

Q Column

Like the other IEPs the *P. mobilis* IEP has a very high pI value of 9.82 and should therefore not bind to a Q column at pH8. However, it was thought that the association with nucleic acids would alter the properties of the protein allowing it to bind to the column as is seen with the other IEP mentioned in this report. It was decided that a Q column could be used, as with *T. carboxydivorans* IEP, to separate the nucleic acids from the IEP. The cell pellet from a 250ml expressed culture was re-suspended in 5ml Q-bind buffer, sonicated and centrifuged. After filtering, the sample was loaded onto a 5ml Q column and the flow-through collected. The proteins were eluted from the column with a gradient of 0.025-2M NaCl. Fractions from the protein peaks seen on the purification trace (Figure 4.59) were electrophoresed on an agarose gel to assess where the nucleic acids had eluted (Figure 4.60).

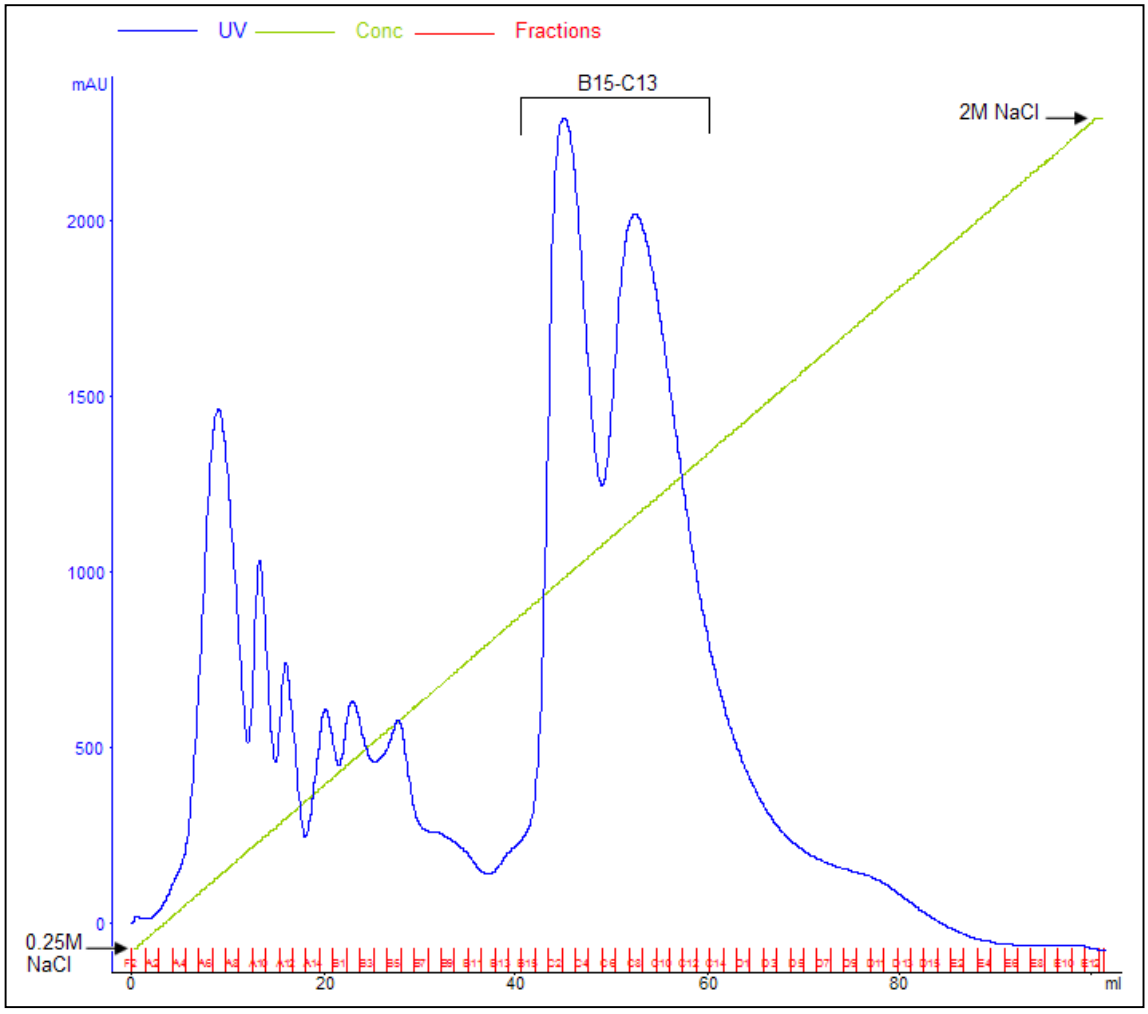

Figure 4.59: A purification trace showing the elution of the soluble fraction from an IEP expression culture from a 5ml Q column.

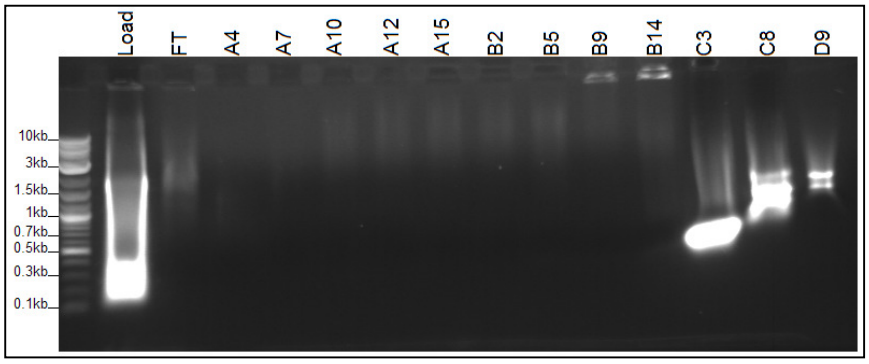

Figure 4.60: A 1%(w/v) agarose gel showing nucleic acid contamination in peak protein fractions from the elution of the IEP expression sample from a Q column.

When loading all the peak fractions onto a protein gel it became apparent that the *P. mobilis* IEP eluted with the fractions that contain the nucleic acids (Figure 4.61). This IEP eluted very broadly and was seen at low levels across fractions B15-C13.

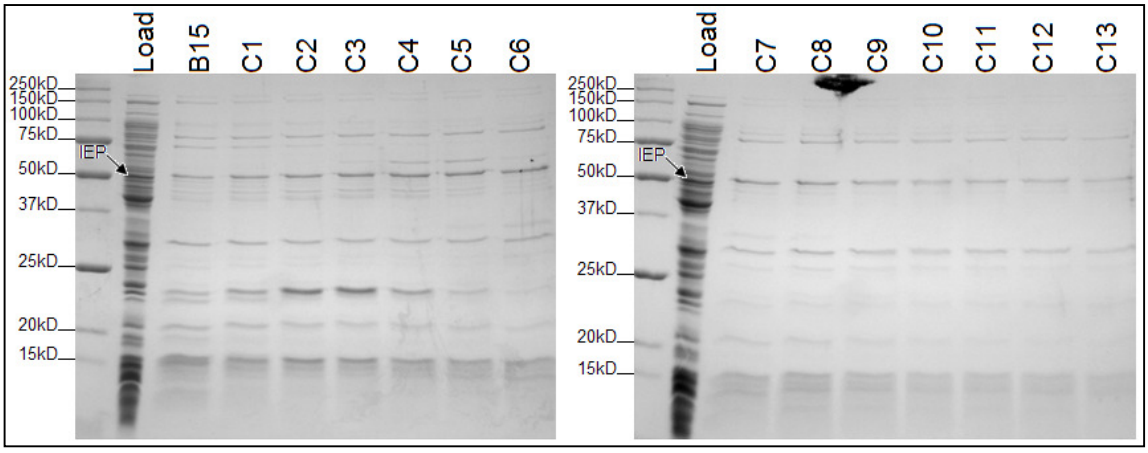

Figure 4.61: 12.5%(v/v) SDS PAGE gel showing fractions from the Q column that contain the *P. mobilis* IEP.

These fractions were assayed for activity using 0.25µl of the fraction as the RT enzyme in a cDNA synthesis reaction using MS2 RNA as template. The PCR used to amplify this cDNA revealed that activity could be found within all these fractions (Figure 4.62). The IEP therefore did successfully bind and elute from the Q column; however, there was a high concentration of nucleic acids contaminating the sample.

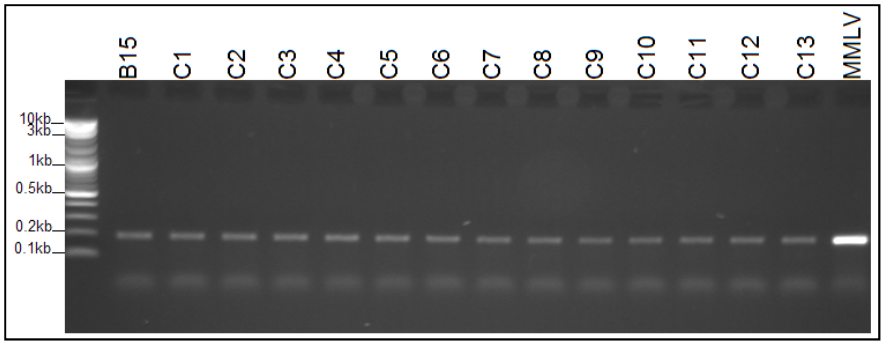

Figure 4.62: A 2%(w/v) agarose gel showing RT activity in the IEP purification fractions from a Q column.

No further attempts were made to purify the *P. mobilis* IEP; however, the enzyme did exhibit activity showing its potential use commercially as an RT enzyme.

B. stearothermophilus IEP

Attempts were made to over-express the IEP from the New Zealand *B. stearothermophilus* using both TB and LB media. However, despite the fact that both media could be used on a 10ml scale, no over-expression of the correct size protein product was seen in either the soluble or insoluble fraction. The expression was repeated several times reducing the culture size to 250ml; however, no over-expression of the IEP on a large-scale was achieved. Therefore, no purification was attempted on this protein.

4.4 – DISCUSSION

This chapter outlined the attempts to purify previously uncharacterised IEPs from thermophilic bacteria. Although not all attempts were successful, RT assays on partially purified enzymes showed that the IEPs from *B. caldovelox*, *T. carboxydivorans* and *P. mobilis* all showed activities with cDNA being created from an MS2 RNA template.

B. caldovelox IEP

Initially the purification of the IEP from *B. caldovelox* proved to be very difficult. Despite the thermophilic nature of the organism from which this IEP was cloned the IEP itself did not appear to display thermostable properties in un-fractionated cell extracts. The soluble protein fraction could not be heated for 5min at 50ºC, in an attempt to reduce background *E. coli* proteins, without the IEP becoming insoluble.

Additional attempts at purification involved utilising an N-terminal his-tag artificially fused to the protein; however, the IEP would not bind to a Ni^{2+} charged column. It was then decided to take advantage of the high pI value of the protein and purify the IEP using an SP column. At pH6.8 the majority of the *E. coli* proteins were not expected to bind the column allowing the IEP to be purified with an NaCl gradient. Unexpectedly, the IEP did not bind to this column and was present in the flow-through. As an alternative method a heparin column was selected to utilise the nucleic acid binding properties of the enzyme; however, once again, the IEP did not bind to the column

It was then found that the protein could bind to an anion exchange Q column even though at pH8 the IEP should have been completely positively charged. The elution of this protein from the Q column revealed it to contain a high level of contaminating nucleic acids and it was hypothesised that the IEP was binding the nucleic acids with very high affinity, preventing it from binding to a heparin column and altering the charge so that it could not bind to the SP column.

PEI was selected as a method to remove nucleic acids, but it was found that a high concentration of NaCl was required to prevent the IEP from precipitating with the nucleic acids. Fortunately, once the nucleic acids were removed, this high NaCl concentration did not interfere with the binding to a heparin column and the IEP could be eluted at approximately 1.6M NaCl. It was also shown that, after nucleic acids were removed, the IEP would bind tightly to the SP column as would be expected with the high pI value of the protein. However, the SP column was not used as a purification step. The preferred method that yielded pure IEP was therefore removal of nucleic acids with PEI at a relatively high NaCl concentration followed by binding and elution, with an NaCl gradient, from a heparin column.

Fractions of the eluted IEP did show RT activity when using MS2 RNA as the template, allowing the potential characterisation of this enzyme as a thermostable RT. The fractions were concentrated and stored in enzyme storage buffer which yield 1.5ml of pure protein, from a 250ml TB expression culture, at 6.4mg/ml.

Sac7d Fusion Proteins

Like the wild-type IEP, the Sac7d fusion proteins could not be heat treated as a method to reduce background proteins without becoming insoluble. Since the

B. caldovelox IEP could only be purified once nucleic acids were removed the same nucleic acids precipitation approach was adopted for the two Sac7d fusion proteins, IEP-Sac7d and Sac7d-IEP.

The IEP-Sac7d fusion behaved very similarly to the wild-type IEP eluting at a high NaCl concentration in fractions free of other contaminating proteins. Once pooled and concentrated into enzyme storage buffer, a 250ml expression sample yielded approximately 1.5ml of IEP-Sac7d at an estimated concentration of 3mg/ml.

The Sac7d-IEP proved more difficult to purify. The fusion product behaved the same as the wild-type IEP and the IEP-Sac7d protein; however, when it eluted from the heparin column, it was found to co-purify with an additional protein. This co-purifying contaminant could not be separated by running an additional SP column. It was expected that the product was a result of cleavage of the Sac7d domain as the contaminating protein was approximately the same size as the wild-type IEP and the product was not seen in the IEP-Sac7d purification. Attempts were made to prevent this cleavage both by using protease inhibitors in the lysis buffers and the purification buffers and by removing the linker domain from the enzyme. Despite these efforts the fusion protein was continually co-purified with this contaminating protein.

In order to obtain pure protein future work would need to involve investigating whether the contamination was in fact due to cleavage of the Sac7d domain. Loading enough of the fractions on a higher percentage protein gel would possibly reveal the presence of a low percentage of the cleaved Sac7d domain running at 7kD. Alternatively the contaminating fragment could be excised from a gel to be sent off for mass spectrometric analysis to give accurate Mr values.

Mass spectrometry analysis could also be carried out after a after a trypsin digest to give an indication of the composition of the contaminating protein. If the contaminating protein was in fact the result of proteolytic cleavage then additional attempts could be made to reduce this occurrence for example by increasing the length of the linker, as the junction of the C-terminus of Sac7d and the N-terminus of the IEP could be providing a proteolytic cleavage site.

Alternatively, the additional product could be due to the presence of another start codon further downstream, possible at the start of the IEP domain. Low level translation could be occurring producing the wild-type IEP lacking the Sac7d domain. This could be eliminated by removal of the natural start codon within the IEP allowing a single read through to be produced.

T. carboxydivorans IEP

Despite the fact that *T. carboxydivorans* grows at an optimum temperature of 55ºC, the IEP in un-fractionated cell extracts could not be heat treated at 50ºC without becoming insoluble, so the background level of *E. coli* proteins could not be reduced.

Since nucleic acids proved to be a problem with the purification of the *B. caldovelox* IEP, PEI precipitation was also adopted as a method to remove the nucleic acids from the sample. However, this IEP was more sensitive to PEI precipitation and was precipitated even at high concentrations of NaCl. Attempts were made to PEI precipitate the nucleic acids in various different NaCl concentrations before binding to a heparin column, but all attempts were unsuccessful. The reasons for failure were most likely either due to all the IEP precipitating with the PEI or due to the high NaCl concentration preventing any of the un-precipitated IEP from being bound to the column. Several alternative columns including Ni^{2+} charged cellulose, SP, Q, heparin and CHT were

adopted and were also used in combined purification steps but no methods proved successful in obtaining pure protein.

It is possible that the *T. carboxydivorans* IEP is not stable in the absence of nucleic acids. Attempts to remove the nucleic acids also precipitated the protein and in some situations, where the nucleic acids were completely removed, the IEP within the fractions quickly precipitated out.

P. mobilis IEP

As is seen for the other IEP outlined in this chapter, the IEP from *P. mobilis* also could not be heat treated in cell extracts in an attempt to reduce background proteins as an initial purification step. Fewer purification methods were attempted on this IEP, than used for *T. carboxydivorans* IEP; however, the same problems were encountered.

Future Attempts

In order to characterise the IEPs as RT enzymes, it was necessary to purify them from proteins that could interfere with the activity, as well as to remove nucleic acids that would contaminate any reactions. This chapter outlined the attempts to purify previously uncharacterised IEPs from thermophilic bacteria. Although not all attempts were successful, RT assays on partially purified enzymes showed that the IEPs from *B. caldovelox*, *T. carboxydivorans* and *P. mobilis* all showed RT activity with cDNA being created from an MS2 RNA template.

It is possible that PEI precipitation could be carried out at higher NaCl in the case of *T. carboxydivorans* and *P. mobilis*. The IEP from *B. caldovelox* eluted from the heparin at a very high NaCl concentration and if the PEI precipitation

was carried out at 1.2-1.4M NaCl it might prevent the IEP from precipitating, but the high affinity to nucleic acids might allow the protein still to bind to the heparin column.

It was also shown in one case that the IEP was associating with RNA rather than DNA and removing the RNA could be achieved using a column with immobilised RNase and therefore preventing any contamination into the sample. However, there is the possibility that removing the RNA could render the protein insoluble, although nucleic acid removal in other cases leaves the IEP in solution.

Purified samples of *B. caldovelox* IEP and the IEP with a C-terminal Sac7d domain were selected for further characterisation of RT activity and will be discussed in the following chapter.

Chapter 5 – Enzyme Characterisation

5.1 – INTRODUCTION

Various assays to characterise the *B. caldovelox* IEP and the Sac7d fusion protein can be carried out and often involve the comparison to retroviral RTs. The backgrounds of these techniques are outlined in the introduction to this chapter.

Quantitative Real Time Polymerase Chain Reaction (qPCR)

qPCR offers a sensitive method for detecting changes in activity of the RTs under differing conditions. The initial first-strand reaction, the RT step, can be performed using a standard thermal cycler allowing accurate incubation time and temperature, followed by a heat denaturation step of the enzyme to prevent any DNA-dependent DNA polymerase activity of the RT from interfering with the qPCR. The cDNA produced by the RT can then be detected using qPCR. Cq values from the qPCR data represent the quantification cycle, also referred to as the cycle threshold, and represents the fluorescence level within the exponential phase of the reaction that is initially detectable above the background. A lower Cq value represents more copies of cDNA being initially present and therefore a more active RT enzyme (Bustin, 2004). Using this sensitive method it is possible to detect subtle changes in RT activity, allowing buffer and reaction temperature optimisation, analysis of enzymatic activity at higher than optimum temperatures, investigation of how the enzyme reacts when given more complex targets, and the thermostability of the enzyme.

Retroviral Activities

The RTs within retroviruses actually have three different activities aiding in the lifecycle of the retrovirus. The RNA genome is synthesised into a DNA strand by the RNA-dependent DNA polymerase activities of the enzyme. The RNase H domain of these RTs is then responsible for degradation of the RNA in the RNA:DNA hybrid, allowing the DNA-dependent DNA polymerase activities of the RT to create dsDNA (as reviewed by Herschhorn and Hizi, 2010). It is known from previous chapters of this report that the IEP shows RNA-dependent DNA polymerase activity and the lack of an RNase H domain presumably prevents the enzyme from degrading the RNA in a RNA:DNA duplex. However, it is not known whether the IEP, like the retroviral RTs, will exhibit DNA-dependent DNA polymerase activity. This activity can be detected using an assay involving a ssDNA template and a FAM (6-carboxy-fluorescein) labelled primer (Wang *et al*. 2004). If an extension event occurs from the FAM-labelled primer, these larger extension products can be detected using an ABI3100 sequencer and GeneScan software. Any products detected that are larger than the 23bp FAM-labelled primer are indicative of DNA-dependent DNA polymerase activity.

Processivity

Processivity of DNA polymerases has been enhanced by the fusion of DNA-binding proteins, such as Sso7d and Sac7d, to either the N- or C-terminus of the proteins. Using the extension assay mentioned above, it was shown by Wang *et al*. (2004) that the presence of an N-terminal Sso7d domain on *Taq* DNA polymerase increases the incorporation of nucleotides, in a single enzymatic event, from 22 to 104 nucleotides. Similarly, the incorporation of Sso7d on the C-terminus of *Pfu* DNA polymerases enhances the processivity of this enzyme to incorporate 6 nucleotides for wild-type *Pfu* to 55 nucleotides with the Sso7d domain. The presence of this DNA-binding domain therefore significantly enhances the processivity of the enzyme. This idea has not yet been adapted for retroviral RTs. In order to test whether Sac7d could have the same effect on

the *B. caldovelox* IEP, the extension assay can be adopted and optimised to allow the measurement of a single enzymatic event. This involves a dilution of the enzyme and a reduction in the reaction time significantly enough to ensure that a single enzymatic event is recorded. As above, a FAM-labelled primer can be used. However, an RNA template would be required to measure the processivity of the RNA-dependent DNA polymerase activity.

Alternatively, a more basic method can be used to measure the processivity. Incorporation of more nucleotides in a single enzymatic event can lead to either shorter extension times required to synthesise a set length of cDNA or to a longer cDNA fragment being created. By recording the time required to synthesise cDNA of specific lengths and comparing wild-type IEP to IEP-Sac7d it would be possible to detect whether the Sac7d is having an influence on the processivity of the enzyme.

Fidelity

DNA-dependent DNA polymerase fidelity assays in the past were often based on the *lac*I system. This blue/white screening system (Brown, 2001) relies on the presence of a fully functional *lac*Z gene coding for β-galactosidase. This is achieved by complementation of two *lacZ* fragments, one from the host strain and the other from a vector. The *E. coli* host strain, TOP10, has the genotype lacZΔM15, which allows complementation with *lac*Zα provided by the pUC vector. The β-galactosidase gene's normal role is to cleave lactose; however, it will also cleave X-Gal producing a blue pigment. When the *lac*I gene is present the cells have a white LacI^{+} phenotype in the presence of X-Gal. The Lac repressor (LacI) will bind the lac operon, blocking the RNA polymerase from transcribing *lac*Z mRNA and therefore preventing production of β-galactosidase. Mutation of *lac*I^{q} by the polymerase can lead to the blue LacI^{-} phenotype, where β-galactosidase transcription is not blocked and colonies will cleave X-Gal

producing blue colonies. The *lac*I gene has been mapped and it is known that, of the 1080bp within the gene, 349 single base substitutions at 179 codons will cause a phenotypic change of white to blue colonies on X-Gal (Provost 1993, cited by Frey 1995).

The error rate (f) of the DNA polymerase can be measured by calculating the number of DNA duplications (d) during the PCR, the fraction of white colonies (F) and with the knowledge that there are 349 phenotypically identified mutations that can occur in the 1080bp *lac*I. The error rate can be calculated using these figures and the equation from Frey *et al*. (1995):

$$f = (-\ln F / d) \times 349$$

This system could be adapted to give an idea of the accuracy of RTs. An RNA strand of *lac*I^{q} can be created with T7 RNA polymerase. However, this step will itself incorporate errors into the *lac*I^{q}. Providing the same template is used for all RTs, this error rate should be constant and therefore not influence the difference in ratio of the white colonies seen with the different RTs. A PCR step would also be necessary, after the cDNA synthesis, to allow *lac*I^{q} ligation into a vector. It would be necessary to ensure that the concentration of cDNA template provided by each RT was identical. This would mean that the duplication rate for each template will be consistent therefore reducing an influence on the error rate by the DNA polymerase. Cycle numbers were kept low and a high-fidelity DNA polymerase used to further reduce any errors produce by the PCR step.

The ligation step can also add an extra problem; if the ligation was unsuccessful, colonies could still grow on the antibiotic medium completely lacking the *lac*I gene and therefore produce a background of false blue colonies. In order to prevent this, a selection technique can be incorporated so that only

colonies positive for *lacI*^{q} would be present on the agar plates. A sucrose lethality system was set up where the *Bacillus subtilis sacB* gene was incorporated into a pUC vector. The *sacB* gene encodes levansucrase that, when expressed by *B. subtilis*, is normally secreted in the culture medium after induction by sucrose (Gay *et al*. 1985). Levansucrase is involved in two main physiological reactions: levan synthesis and sucrose hydrolysis. When Gram-negative bacteria containing *sacB* are grown on media containing 5-10% sucrose, levansucrase is expressed and it is believed that it becomes trapped in the periplasmic space (Bramucci and Nagarajan, 1996). Accumulation of levans in the periplasm proves lethal by causing lysis of the cell (Recorbet *et al*. 1993). The fidelity system was designed so that the *lacI*^{q} would be inserted into a *sacB* gene within pUC19 and plated onto sucrose plates. Where *lacI*^{q} failed to ligate into the vector, *sacB* will remain intact resulting in lethality on sucrose plates and therefore no background blue colonies. A positive insertion of *lacI*^{q} into *sacB* interrupts this gene, preventing levansucrase expression and therefore preventing lethality on sucrose. This system will ensure that only colonies positive for the *lacI*^{q} will be present on the agar plates.

5.2 – MATERIALS AND METHODS

Nuclease Assay

An optimum enzyme concentration was added to 1x RT buffer and 500ng of either MS2 RNA, lambda DNA (Fermentas, York, UK) or pET24a vector DNA and the volume made up to 15µl with nuclease-free water. The reaction was incubated for 10h at 37ºC and then the entire sample electrophoresed on an agarose gel to assess for nuclease activity. A control was run in parallel containing nucleic acids but no enzyme.

qPCR

This reaction used cDNA, synthesised by the RT enzyme, as a template for the PCR. The final reaction contained 1x MESA Green [no ROX] Mastermix (Eurogenetic, Belgium), 15pmol of primers MS2:3231_F and MS2:3395_R (Appendix I), 1µl of the cDNA synthesis reaction and the volume made to 50µl with nuclease-free water. A LightCycler®480 (Roche, Welwyn Garden City, UK) was used with the following cycling conditions:

94ºC 4min

95ºC 3s
60ºC 1min } 45 cycles

The cycling conditions were then directly followed with a melt curve analysis from 60-95ºC at 0.04ºC/s.

Reaction Buffer optimisation

A proprietary RT buffer (GeneSys Ltd, Camberley, UK) was altered, one component at a time, to analyse the best concentrations required for optimum cDNA synthesis by the IEP and IEP-Sac7d. The following components were tested during the cDNA synthesis step with the following final concentrations:

Tris: pH 7.9, 8.0, 8.1, 8.2, 8.3, 8.4, 8.5, 8.6, 8.7, 8.8. 8.9, 9.0

Ammonium sulphate: 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50mM

KCl: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100mM

K_2SO_4: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100mM

$MgSO_4$: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10mM

All reactions were performed in duplicates and 1µl of the above reaction was used as template in a qPCR and the protocol followed as above.

Optimum Reaction Temperature and High Reaction Temperatures

The final cDNA synthesis reaction contained 1x optimum RT reaction buffer, 0.5mM dNTPs, 15pmol MS2:3395_R, 1ng MS2 RNA, optimum RT enzyme concentration, and the final volume made up to 20µl with nuclease-free water. Using a Veriti® thermal cycler (Applied BioSystems, UK) samples were incubated for 30min at various set temperatures. The temperatures varied from 42-76ºC in increments of 2ºC and were followed by a heat denaturation step at 95ºC for 20min. All reactions were performed in duplicate. 1µl of this reaction was used as a cDNA template in a qPCR as mentioned earlier.

Metal Ion Assay

cDNA Synthesis Step

The cDNA synthesis reaction contained 1x optimum buffer (with no $MgSO_4$), 0.2mM dNTPs, 15pmol MS2:3395_R, 20ng MS2 RNA, optimum RT enzyme concentration, 0-10mM divalent cation and the final volume made to 20µl with nuclease-free water. The reaction was incubated at the optimum enzyme temperature for 30min and then heated to 95ºC for 20min to irreversibly denature the RT enzyme.

PCR Step

2µl of the above reaction was used as template in a PCR that also contained 1x *Taq* mastermix, 25pmol MS2:3231_F, 25pmol MS2:3395_R and the volume made up to 50µl with nuclease-free water. The reaction was cycled as follows:

95ºC	3min	
95ºC	10s	} 30cycles
55ºC	10s	
72ºC	20s	
72ºC	7min	

10µl of the reaction was analysed on an 2%(w/v) agarose gel.

Thermostability Assay

Temperature Incubation

4 different reactions were set up containing 1x optimum reaction buffer (with no $MgSO_4$), 0.5mM dNTPs and RT enzyme. Once the basic mastermix was set up, the components were added to the four individual reactions as follows:

1) 2mM $MgSO_4$
2) 2mM $MgSO_4$ and 1ng MS2 RNA
3) 2mM $MgSO_4$ and 1µg yeast tRNA
4) 1ng MS2 RNA

The reactions did not contain any MS2:3395_R and therefore no cDNA would be synthesised during the incubation steps. The reactions were heated to 60ºC for time intervals of 0, 5, 10, 15, 20, 25, 30, 40, 50, 60, and 90min. All reactions were performed in duplicate.

cDNA Synthesis Step

Following the incubations as described above, 15pmol MS2:3395_R and, if necessary, 1ng MS2 RNA and 2mM $MgS0_4$ were added to each reaction and incubated for 30min at the optimum enzyme reaction temperature, followed by a denaturation step of 95ºC for 20min. 1µl of this template was then used as the DNA template in a qPCR as mentioned above.

Fidelity Assay – Set Up and Proof of Principle

Creating pCR®2.1 $lacI^q$ - Gene Amplification

lacI^q was amplified from pTTQ18NHK vector (GeneSys Ltd, Camberley, UK) using PCR. The reaction contained 1x HF buffer, 0.2mM dNTPs, 25pmol of primers *lacI*^{q}_F_SacII and *lacI*^{q}_R_SacII (Appendix I), 10ng pTTQ18NHK, 1U

Phusion® DNA polymerase and made up to 50µl with nuclease-free water. The reaction was cycled as follows:

98ºC	30s		
98ºC	10s	}	20 cycles
55ºC	10s		
72ºC	1min		
72ºC	7min		

The reaction was purified using the Promega Wizard® SV Gel and PCR Clean-up system and eluted in 50µl nuclease-free water.

Creating pCR®2.1 lacI^q – A-Tailing

7µl of the purified PCR was added to 1x *Taq* buffer (GeneSys Ltd, Camberley, UK), 0.2mM dATP (Roche, Welwyn Garden City, UK), 5U *Taq* DNA polymerase and the final volume made up to 10µl with nuclease-free water. The reaction was incubated at 70ºC for 15-30min.

Creating pCR®2.1 LacIq – General Methods

The protocols followed to allow TA cloning and sequencing of the pCR®2.1 vector can all be found in Chapter 2 methods.

Creating pUC19 sacB – Gene Amplification

sacB was amplified from pET24a_*sacB* (GeneSys Ltd, Camberley, Surrey) in a PCR containing 1x HF buffer, 0.2mM dNTPs, 25pmol of primers *sacB*_F_*Aat*II and *sacB*_R_*Aat*II (Appendix I), 1U Phusion® DNA polymerase and the final volume made to 50µl with nuclease-free water.

The reaction was cycled as follows:

98ºC	30s	
98ºC	10s	20 cycles
55ºC	10s	
72ºC	1min 45s	
72ºC	7min	

The reaction was cleaned up using Promega Wizard® SV Gel and PCR Clean-up system and DNA eluted in 50µl nuclease-free water.

Creating pUC19_sacB - General Methods

The protocols followed to allow the ligation of the *sacB* gene into pUC19 and the sequencing of the construct are all detailed in Chapter 2 methods.

Proof of Principle

PCRs were set up using three different polymerases, *Taq*, *Taq*:*Pfu* (20:1) and Phusion®. The reactions contained 1x recommended reaction buffer, 0.2mM dNTPs, 50pmol *lacI*^{q}_F_SacII, 50pmol *lacI*^{q}_R_SacII, 5ng pCR®2.1_*lacI*^{q}, optimum DNA polymerase concentration and the volume made up to 50µl with nuclease-free water. The reactions were cycled as follows

Taq/*Taq*:*Pfu*			Phusion®		
95ºC	3min		98ºC	30s	
95ºC	10s	18 cycles	98ºC	10s	18 cycles
55ºC	10s		55ºC	10s	
72ºC	1min 20s		72ºC	50s	
72ºC	7min		72ºC	7min	

A 5µl aliquot of each reaction was electrophoresed on an agarose gel with 0.5µg 2-log ladder marker (NEB, Hitchin, UK) to allow an estimation of the yield.

The *lacI*^{q} amplified gene was purified, digested with *Sac* II and ligated into *Sac* II digested pUC19U_*sacB* following protocols as detailed in chapter 2. The construct was then transformed into electrocompetent TOP10 (Appendix II) and spread onto LB plates containing 80µg/ml methicillin, 20µg/ml ampicillin, 40µg/ml X-Gal and 10%(w/v) sucrose. After incubation for 24h at 37ºC the number of blue and white colonies could be counted.

RT Fidelity Assay

Generating lacI^{q} RNA

PCR®2.1_*lacI*^{q} was linearised in a RE reaction containing 1x NEB RE buffer 4, 10U *Bgl* II (NEB, Hitchin, UK), 1µg pCR®2.1_*lacI*^{q} and the volume made to 100µl with nuclease-free water. The reaction was incubated for 1h at 37ºC and then purified using Promega Wizard® SV Gel and PCR Clean-up system with the DNA eluted in 50µl nuclease-free water.

AmpliScribe™T7 High Yield Transcription Kit (Cambio Ltd, Cambridge, UK) was used to generate RNA from pCR®2.1_*lacI*^{q} plasmid. The reaction components were added in the following order: RNase-free water, 100ng linearised pCR®2.1_*lacI*^{q}, 1x AmpliScribe T7 reaction buffer, 7.5mM ATP, CTP, GTP, UTP, 10mM DTT, 0.5µl RiboGuard RNase Inhibitor, and 2µl Ampliscribe T7 Enzyme solution, with the final reaction volume equalling 20µl. The reaction was then incubated for 2h at 37ºC. 1µl (1MBU) RNase-free DNase I was added, incubated for 15min at 37ºC and then heat-inactivated at 75ºC for 10min.

cDNA Synthesis Step

Four separate first-strand reactions were set up, each for a different RT enzyme. Each reaction contained 1x optimum RT buffer, 0.5mM dNTPs, 15pmol *Lacl*^{q}_F_SacII, *lacl*^{q} template RNA (estimated at 50ng), optimised RT enzyme concentration and the volume made to 20µl with nuclease-free water. The reactions were incubated for 30min at the optimum enzyme reaction temperature followed by a denaturation step at 95ºC for 20min

PCR step

Initially, dilutions of the cDNA synthesis reaction were set up and used in a PCR containing 1x HF buffer, 0.2mM dNTPs, 25pmol *Lacl*^{q}_F_SacII and *Lacl*^{q}_R_SacII, cDNA synthesis reaction as template, and 1U Phusion® DNA polymerase and the volume made to 50µl with nuclease-free water. The reaction was cycled as follows:

98ºC	30s		
98ºC	10s	}	18 cycles
55ºC	10s		
72ºC	50s		
72ºC	7min		

Once the same concentration of starting template had been determined, the reaction was repeated to ensure the same amount of cDNA had been amplified.

Cloning Step

The protocols followed to allow the purification, digestion and ligation into pUC19_*sacB* are detailed in Chapter 2. The ligated constructs were transformed into electrocompetent TOP10 and plated onto LB with methicillin, ampicillin, X-Gal and 10%(w/v) sucrose. After 24h incubation at 37ºC the white and blue colonies could be counted.

DNA-Dependent DNA Polymerase Assay

The FAM-labelled primer, -40M13LFF (Wang *et al*. 2004 and Appendix I) was initially annealed to the template in a reaction buffer that contained 1x optimum reaction buffer, 0.25mM dNTPs, 1x BSA, 100nM M13mp8 template, 50nM -40M13LFF, and the volume made to 18µl with nuclease-free water. The reaction buffer was placed into a glass beaker with 95ºC water and allowed to cool to room temperature.

Taq DNA polymerase was used as a control reaction with 0.1U of the polymerase added to the reaction, incubated at 72ºC for 5min and the reaction stopped with the addition of EDTA to a final concentration of 20mM.

For the RT experiment, the optimum enzyme concentration was added and the reaction incubated at the RT optimum reaction time for 30min before being stopped with the addition of EDTA to the final concentration of 20mM.

The reactions were ethanol-precipitated with the addition of 10µl 125mM EDTA, and 112µl of 100%(v/v) ethanol, and incubated at -20ºC for 20min. After incubation the reactions were centrifuged at 14,000xg for 20min, the supernatant discarded and the pellet washed with 100µl 70%(v/v) ethanol. The ethanol wash was discarded and the pellets dried using a vacuum centrifuge. The pellets were re-suspended in 10µl ABI Hi-Di™ formamide. A second buffer was made up to contain 4µl GeneScan™350 ROX™ size standard and 140µl of fresh formamide. 9µl was dispensed into wells of a sequencing plate along with 1µl of the reaction in Hi-Di™ formamide and the plate gently mixed.

The assay reaction was analysed using the GeneScan programme of an ABI3100 sequencer.

RNA-Dependent DNA Polymerase Activity – Processivity Assay

This experiment measures the extension event occurring from a FAM-labelled primer, FAM_MS2:3193R (Appendix I), annealed to MS2 RNA, by RNA-dependent DNA polymerase activity. The basics of this method were the same as those for the above assay; however, several adjustments were made that will be detailed in the results section.

Basic Processivity Assay

cDNA Synthesis Step

The final reaction contained 1x optimum reaction buffer, 0.5mM dNTPs, 15pmol MS2:3339_R, 20ng MS2 RNA, and an optimum enzyme concentration, and the volume made up to 20µl with nuclease-free water. Samples were incubated at the enzyme optimum reaction temperature for 0, 1, 2, 3, 4, 6, 8, 10, 12, 14, 16, 18, 20, 25, 30min and then immediately incubated at 95ºC for 20min. The reactions then had an additional 20µl of nuclease-free water added.

PCR step

The final reaction contained 1x *Taq:Pfu* (20:1) mastermix, 15pmol MS2:3395_R and 15pmol of either MS2:13_F (Appendix I) or MS2:1868_F (Appendix I), and 1µl of the above reaction as template and the volume made to 25µl with nuclease-free water.

The reaction was cycled as follows

94ºC 3min

94ºC 10s
55ºC 10s
72ºC 1min/kb
} 20 cycles

72ºC 7min

5µl of the reaction was then analysed on an agarose gel.

Complex Target Assay

cDNA Synthesis

The reaction used either random nonamers, Oligo (dT) or specific primers to amplify targets from total human placental RNA (Ambion, Warrington, UK). The specific primers were designed to create cDNA from 28S rRNA, ATP Synthase, glyceraldehydes-3P dehydrogenase (GAPDH) and β-2-microglobulin mRNAs. The reaction contained 1x optimum RT reaction buffer, 0.5mM dNTPs, 50pmol primer mix, RNA (100, 10, 1, 0.1, 0.01, and 0ng), and an optimum enzyme concentration, and the final volume made up to 20µl with nuclease-free water. The reaction was heated to the optimum RT reaction temperature for 30min followed by an enzyme denaturation step at 95ºC for 20min.

qPCR

1µl of the above template was used in a reaction with 1xMESA [no ROX] Mastermix, 1.5µm specific primer [either 28S, ATP Synthase, GAPDH or β-2-microglobulin (Appendix I), and the volume made to 50µl with nuclease-free water.

A LightCycler®480 was used with the following cycling conditions:

94ºC 4min

95ºC 3s
60ºC 1min } 45 cycles

followed by a melt curve analysis from 60-95ºC at 0.04ºC/s.

5.3 – RESULTS

Purification was not successful for the IEPs from *P. mobilis*, *T. carboxydivorans* and the New Zealand strain of *B. stearothermophilus*; therefore, no characterisation of these enzymes could be performed. All the characterisation was carried out on the IEP from *B. caldovelox* and the same IEP with a C-terminal Sac7d domain, with comparisons to each other where necessary and also, in some assays, comparisons to other RTs.

Optimum Enzyme Level

The purified IEP and IEP-Sac7d were stored in enzyme storage buffer at concentrations of 6.4mg/ml and 3mg/ml, respectively. In order to find the optimum level of enzyme required, an RT-qPCR was performed with a proprietary RT buffer and 1:10 dilutions of the enzyme. Once an optimum level was found using this dilution series, this was repeated with 1:2 dilutions of the enzyme and it was found that, for both IEP, 1ng of enzyme was optimal for a 20µl RT reaction.

Nuclease Assay

Once the working concentration of the enzyme had been optimised, it was necessary to test for any nuclease activity that might inhibit or interfere with the enzyme characterisation. It was important that the samples did not contain any RNase activity as this would degrade the template for the cDNA synthesis step, and DNase activity would interfere with both the cDNA synthesised and any subsequent PCR step. 1ng of the enzyme was incubated at 37ºC for 10h, with RT buffer and 500ng of either MS2 RNA, lambda DNA or pET24a vector DNA

and the reaction electrophoresed on an agarose gel (Figure 5.1). A control with no enzyme present was also run in parallel.

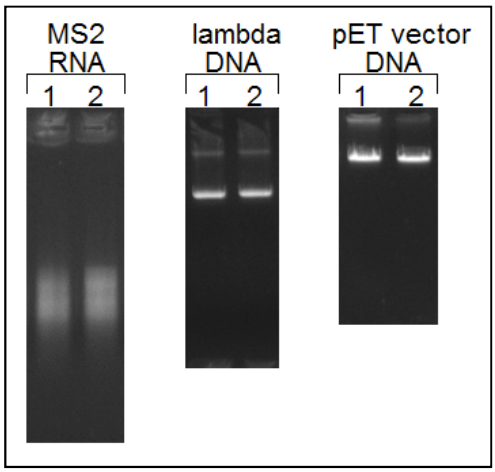

Figure 5.1: A 1.5%(w/v) agarose gel showing the results for the nuclease activity assay with lane 1: 1ng enzyme and lane 2: no-enzyme control (NEC). The results shown are for the wild-type IEP.

The results for both IEP and IEP-Sac7d showed no DNase activity in the enzyme sample compared to the NEC. The results initially looked as if there was slight RNase activity when incubating the RNA with the purified enzyme. However, there is some fluorescence seen in the well of the gel suggesting that the enzyme was binding some of the RNA and preventing it from entering the gel. This was much more apparent when using higher concentrations of enzyme so it was concluded that no nuclease activity was present in the sample.

Reaction Buffer Optimisation

In order to characterise the enzymes it was important that they were being assayed in an optimum reaction buffer allowing their comparison to other optimised RTs. This was performed using a proprietary RT buffer and altering

the concentration of one component at a time. A first-strand reaction was performed followed by qPCR to detect the difference in the enzymatic activity. The optimisation was carried out on both IEP and IEP-Sac7d.

Both enzymes proved to be very broad with respect to their pH preference with only very slight changes in activity detected across the tested pH range. However, it was noted that the Sac7d domain increased the requirement for salts in the form of ammonium sulphate and potassium sulphate. The final optimised 1x reaction buffers were as follows:

	IEP	IEP-Sac7
50mM Tris	pH8.3	pH8.3
Ammonium Sulphate	15mM	25mM
Potassium Sulphate	30mM	80mM
Magnesium Sulphate	2mM	2mM
DMSO	2%	2%

Optimum Reaction Temperature

Once the reaction buffer was optimised, it was also necessary to optimise the reaction temperature. Using an RT step, with MS2 RNA as template, it was possible to vary the temperature of the first-strand reaction and then analyse the enzymes activity using qPCR. This assay was also used to compare the IEPs to MMLV-RT to analyse the difference between a mesophilic RT and an RT isolated from a thermophilic organism.

The first-strand reaction was performed at various temperatures and the qPCR data compared for IEP, IEP-Sac7d and MMLV-RT (Figure 5.2). The Cq values can be seen in Table 5.1.

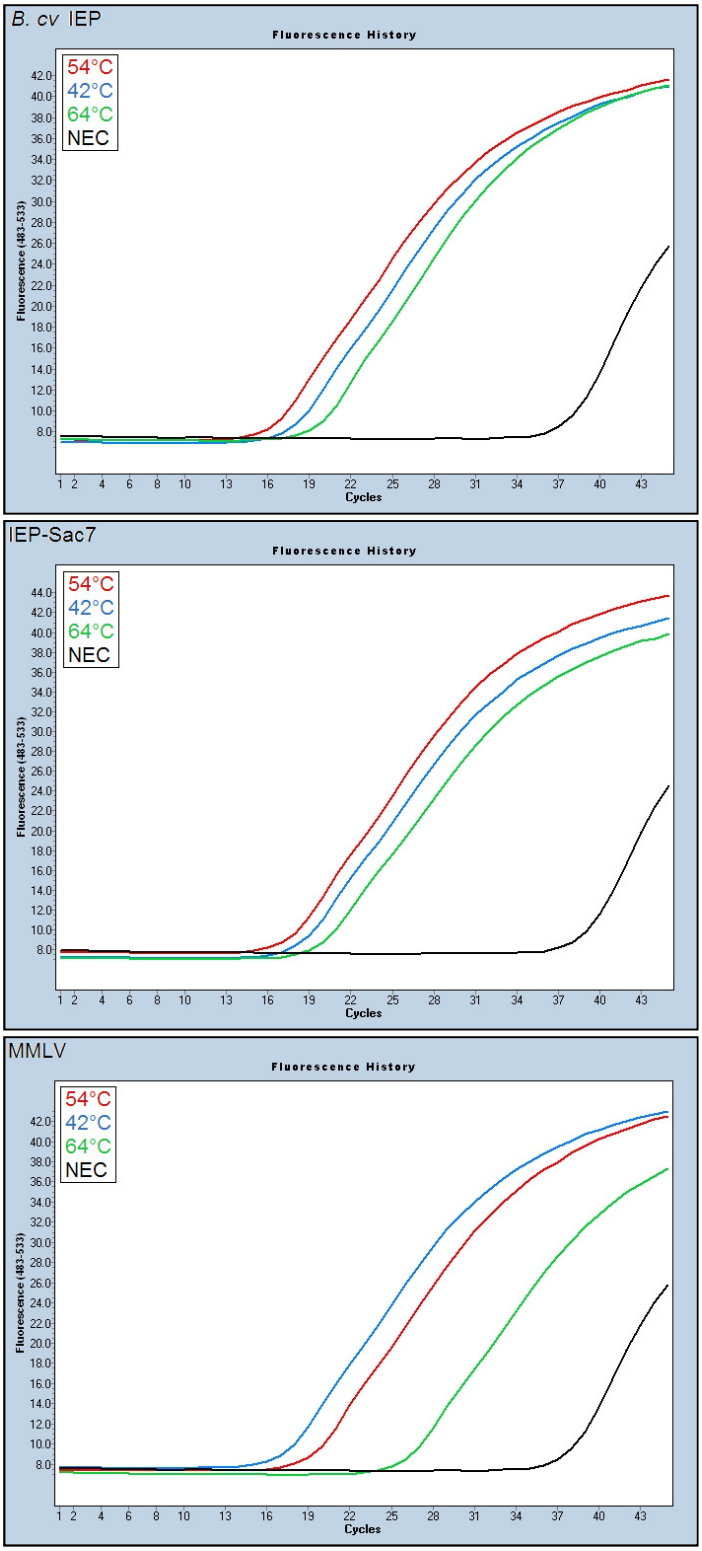

Figure 5.1: qPCR data showing the activity of the IEP, IEP-Sac7d and MMLV-RT at 42ºC(blue), 54ºC(red) and 64ºC (green). The black line corresponds to a no-enzyme control (NEC).

Reaction Temperature (ºC)	IEP (Cq value)	IEP-Sac7d (Cq value)	MMLV-RT (Cq value)
42	16	17	11
54	14	15	16
64	17	19	24
NEC	36	36	36

Table 5.1: A table showing the number of cycles required for the fluorescence to becomes significantly above the background (Cq value) using different RT enzymes at different reaction temperatures.

Both IEPs gave very similar results, putting their optimum reaction temperature at 54 ºC, 12ºC higher than that recommended for MMLV-RT. The IEP had more activity, as seen by an earlier Cq value, at the lower temperature range than at the temperature above its optimum. The Sac7d domain did not appear to have a major influence on the optimum reaction temperature although Cq values were slightly higher at each temperature for IEP-Sac7d. For example at 64ºC the Cq value for the IEP was 17 but was later at 19 cycles for IEP-Sac7d. Comparison of MMLV-RT with the IEPs revealed that increasing the temperature above the optimum has more of a detrimental effect on the retroviral RT than the IEPs. At 64ºC a Cq value of 24 cycles was detected for MMLV-RT, compared to 16 cycles seen when using the IEP.

Higher cDNA Synthesis Reaction Temperatures

An additional test was carried out to measure the activity of the RT enzymes, both MMLV-RT and IEPs, at reaction temperatures higher than their optima. The RT step was performed at increasingly higher temperatures, with MS2 RNA as the template, and the enzyme activity analysed using qPCR (Figure 5.3 and Table 5.2).

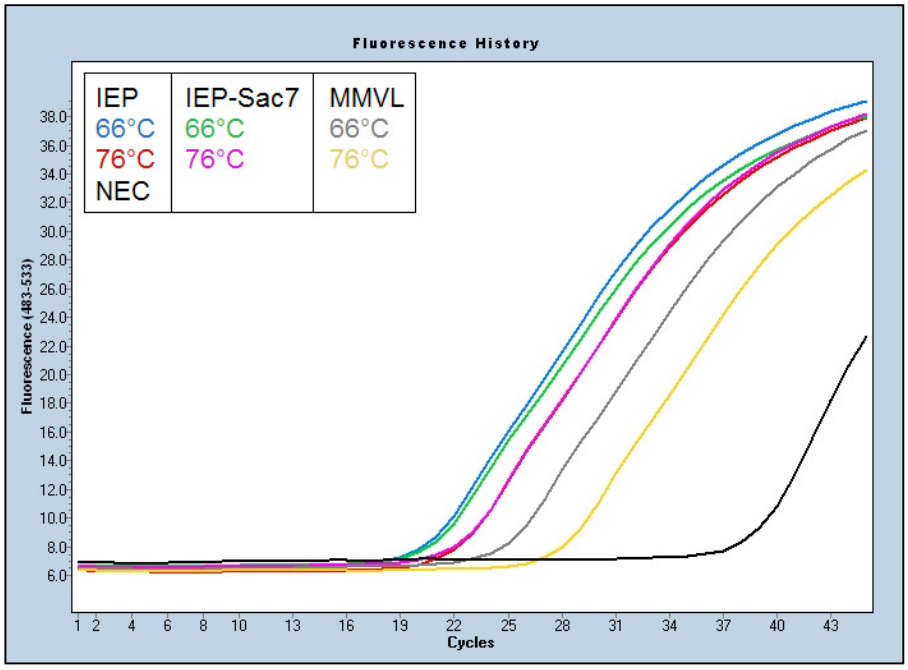

Figure 5.3: qPCR data showing the result of increasing high temperature for cDNA synthesis with IEP, IEP-Sac7d and MMLV-RT. The black line corresponds to an NEC.

Reaction Temperature (ºC)	IEP (Cq value)	IEP-Sac7d (Cq value)	MMLV-RT (Cq value)
66	19	19	23
76	21	21	27
NEC	37	37	37

Table 5.2: A table showing the number of cycles required for the fluorescence to becomes significantly above the background (Cq value) using the different RT enzymes at different reaction temperatures.

The two reaction temperatures shown on these qPCR data are higher than the optimum reaction temperature of all three enzymes. The IEP and IEP-Sac7d both perform very similarly to each other at 66ºC and at 76ºC, with the Sac7d domain appearing to have no effect on the enzymatic activity at these temperatures. MMLV-RT, however, is more affected by the increasing temperature and this can be seen in two ways from these graphs. Firstly, the enzyme is less active at 66ºC than the IEPs at 76ºC. This can be seen by the

measured Cq value of 23 cycles for MMLV-RT at 66ºC compared to a Cq value of 21 cycles with IEP at 76ºC. Secondly, the effect of increasing the temperature is greater with MMLV-RT as a 10ºC rise in reaction temperature results in an increase of 4 cycles required until the fluorescence is detectable above the background compared to an increase of only 2 cycles as seen with the IEPs.

The difference in the two enzymes activities can be seen more clearly on this second qPCR data graph (Figure 5.4).

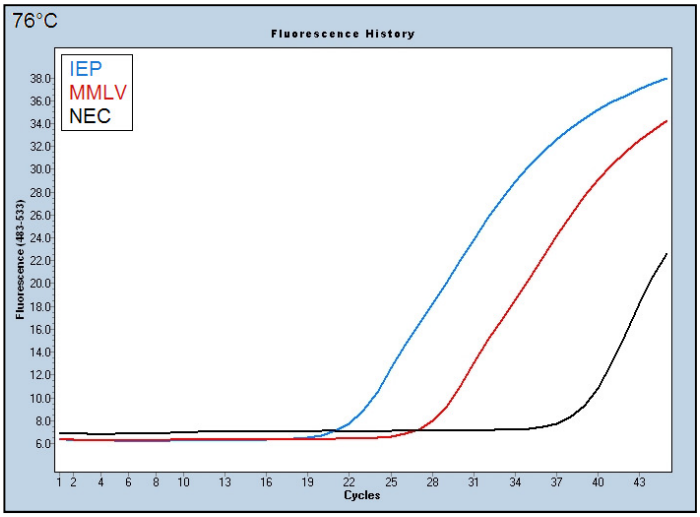

Figure 5.4: qPCR data showing the RT activity of MMLV-RT (red) and IEP (blue) at 76ºC. The black line corresponds to an NEC.

Since IEP and IEP-Sac7d performed the same at this temperature only the wild-type IEP data were included on the graph. These data show a marked difference in the enzymes' activity with a Cq value of 21 cycles for IEP compared to 27 cycles for MMLV-RT, indicating that the IEPs have a much greater RT activity at 76ºC than MMLV-RT.

Ion Usage

Normally, retroviral RTs require either Mn^{2+} or Mg^{2+} ions for optimum activity and an assay was carried out to analyse how broad or specific the IEPs were in terms of their preferred source of divalent cations. The optimum buffer used had $MgSO_4$ removed and this was replaced with increasing concentrations of one of the following:

(a) $MgCl_2$	(b) $CoSO_4$
(c) $MgSO_4$	(d) $CaCl_2$
(e) $MnCl_2$	(f) $NiCl_2$
(g) $ZnSO_4$	(h) $CuCl_2$

2µl of the cDNA synthesis step was then used in a 50µl PCR with 30 cycles performed and 10µl electrophoresed on an agarose gel (Figure 5.5).

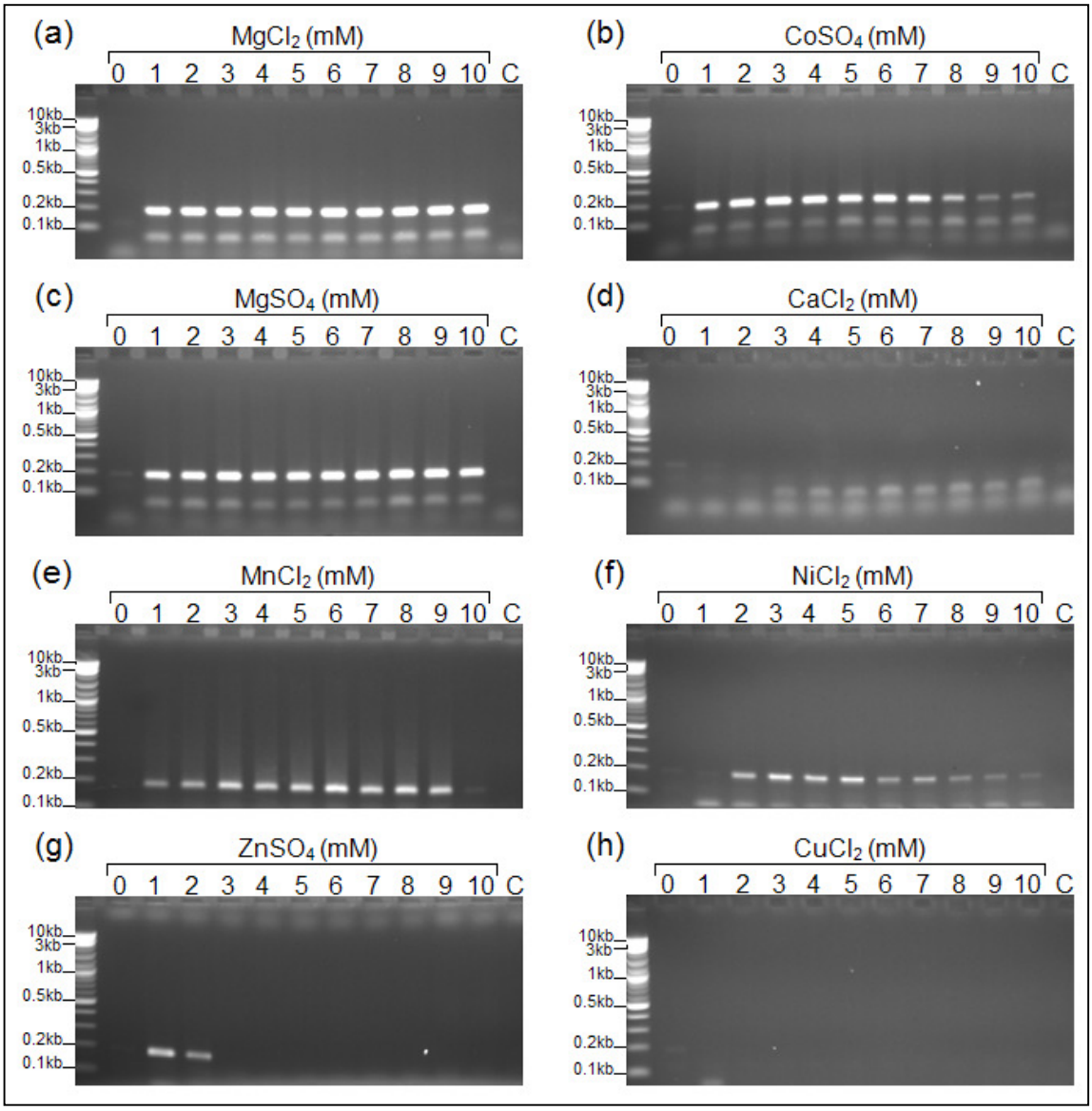

Figure 5.5: A 2.5% agarose gel showing the amplification of cDNA produced by the IEP with increasing concentrations of varying divalent cation sources. C corresponds to an NEC.

$MgCl_2$ and $MgSO_4$ were acting as controls and no variations in enzymatic activity were detected. After Mg^{2+} the RTs next preferential ion was Co^{2+}, at low concentrations, followed by Mn^{2+} at higher concentrations and then, to a lesser extent, the lower concentrations of Ni^{2+}. There was also low level activity

detected when 1-2mM Zn^{2+} was supplied. There did not appear to be any activity when Ca^{2+} and Cu^{2+} were the divalent cation source.

To test whether this was a true reflection of the divalent cations used by the RT, and not a result of inhibition of the PCR step, the 10mM divalent cation cDNA synthesis reaction was added to a known working PCR to amplify a 500bp fragment from lambda DNA (Figure 5.6). The final concentration of each ion in the sample corresponded to 0.4mM, the maximum that would be found in the previous assay.

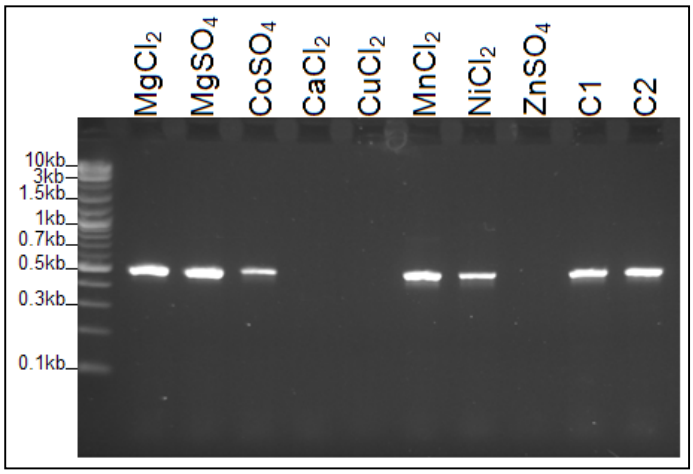

Figure 5.6: A 1%(w/v) agarose gel showing the effect, on a 500bp lambda PCR, of the addition of 0.4mM divalent cations from a first-strand cDNA synthesis reaction. C1 contained cDNA synthesis reaction with 0mM divalent cation and C2 was a standard 500bp PCR sample with no first-strand reaction added.

The two controls showed the same level of amplification, so that it was clear any differences observed in the test was due to ion concentration and not due to interference from a separate component from the cDNA synthesis reaction. 0.4mM $MgCl_2$, $MgSO_4$ and $MnCl_2$ appeared to have no effect on the 500bp lambda PCR. $CoSO_4$ and $NiCl_2$ were showing a partial inhibitory affect, while $CaCl_2$, $CuCl_2$ and $ZnSO_4$ were completely inhibitory to the PCR. Therefore, the

results seen with these ions on the RT step could be due to the inhibitory affect on PCR rather than the RT not being able to utilise these ions. An additional lambda PCR showed that diluting the ions to 0.04mM was enough to prevent complete inhibition of the PCR. The RT ion usage experiment was repeated using those ions that inhibited the PCR with the cDNA synthesis reaction diluted 10-fold before adding to the PCR to ensure that no more than 0.04mM of the ions were carried over to the PCR step (Figure 5.7).

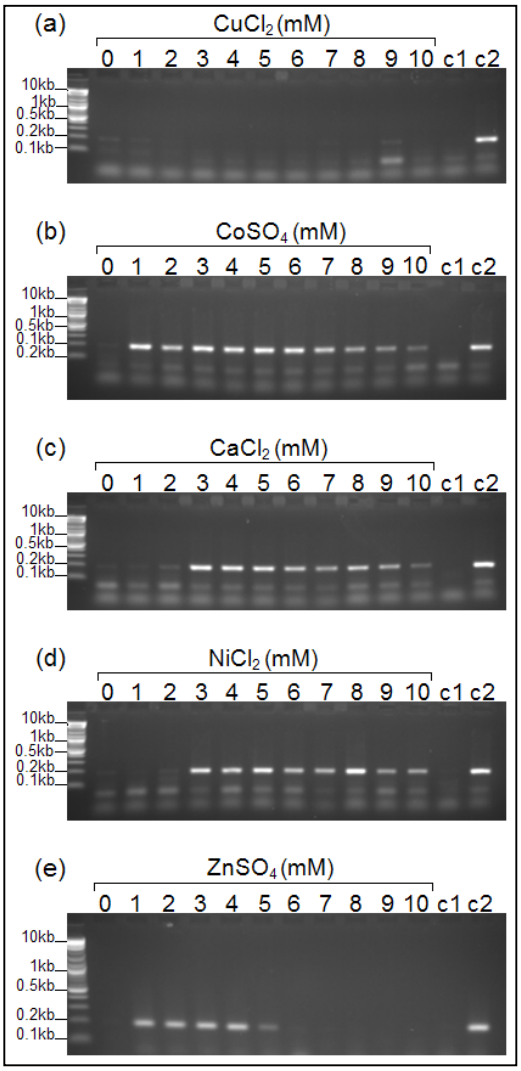

Figure 5.7: A 2.5% agarose gel showing the amplification of cDNA produced by the IEP with increasing concentrations of varying divalent cation sources. c1 corresponds to an NEC and c2 corresponds to a first-strand reaction carried out with $MgSO_4$.

The results of the second ion usage assay showed that the RT could utilise Zn^{2+} at low levels before it presumably becomes inhibitory to the enzyme. Although the IEP is not able to utilise Cu^{2+}, unless specifically at 9mM, it was able to use Co^{2+}, Ca^{2+} and Ni^{2+}, and it is possible that the decrease in activity seen with increasing ion concentration is an artefact due to increasing inhibition in the PCR. While these results appear that this type of RT is very broad in terms of its divalent cation usage, it does not appear to be uncommon for reverse transcriptases. Unpublished data performed in parallel to this experiment revealed MMLV-RT to display a similar pattern while Filler and Lever (1997) reported HIV-1 RT to be capable of utilising a range of different cations with Mg, Mn, Cu and Co all being utilised by the enzyme with differing efficiencies.

Thermostability

To test the thermostability of the IEPs compared to MMLV-RT, the enzymes were heated, in their reaction buffer, at 60ºC for various time intervals before primers were added to allow cDNA synthesis to take place. The cDNA synthesis step was carried out at each enzymes optimum reaction conditions for 30min and then analysed using qPCR.

Initially, the 60ºC incubation step was carried out with no RNA present to avoid the effect of RNA degradation at the higher temperatures (Figure 5.8 and Table 5.3).

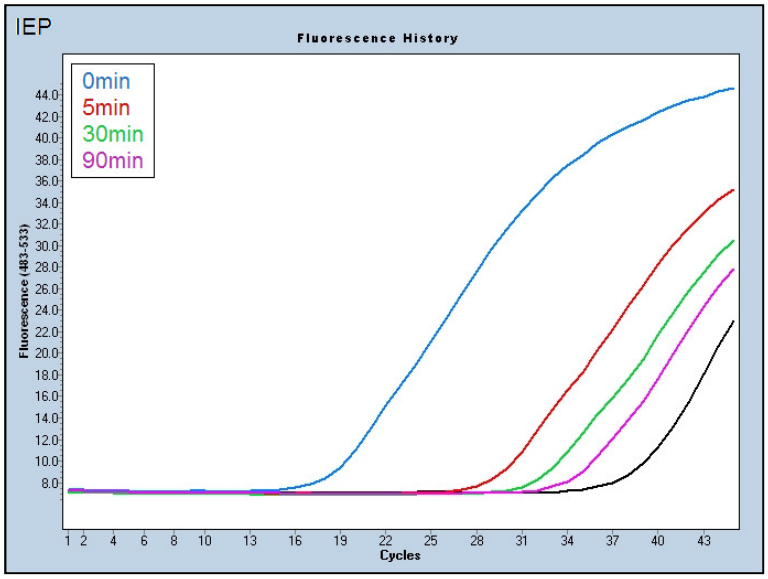

Figure 5.8: qPCR data showing the result of cDNA synthesis once the IEP had been incubated for set time intervals at 60ºC.

Incubation Time (min)	Cq value
0	15
5	27
30	30
90	32
NEC	36

Table 5.3: A table showing the Cq values for an RT reaction with IEP after incubation at set times at 60ºC. No RNA was present in the incubation prior to the cDNA synthesis reaction.

The initial results of thermostability seemed to show a large decrease in enzymatic activity even after just 5min at 60ºC. This was indicated by a decrease in Cq value from 15 cycles with no incubation to 27 cycles after 5min. This was unexpected as the IEP had been isolated from *B. caldovelox* with an optimum growth temperature of 70ºC. Further decrease in activity could be seen with increasing length of incubation at 60ºC.

It was decided to repeat the experiment in the presence of RNA to see if it would act to stabilise the enzyme during the incubation step. The experiment was repeated both with template MS2 RNA as well as yeast tRNA (Figure 5.9 and Table 5.4).

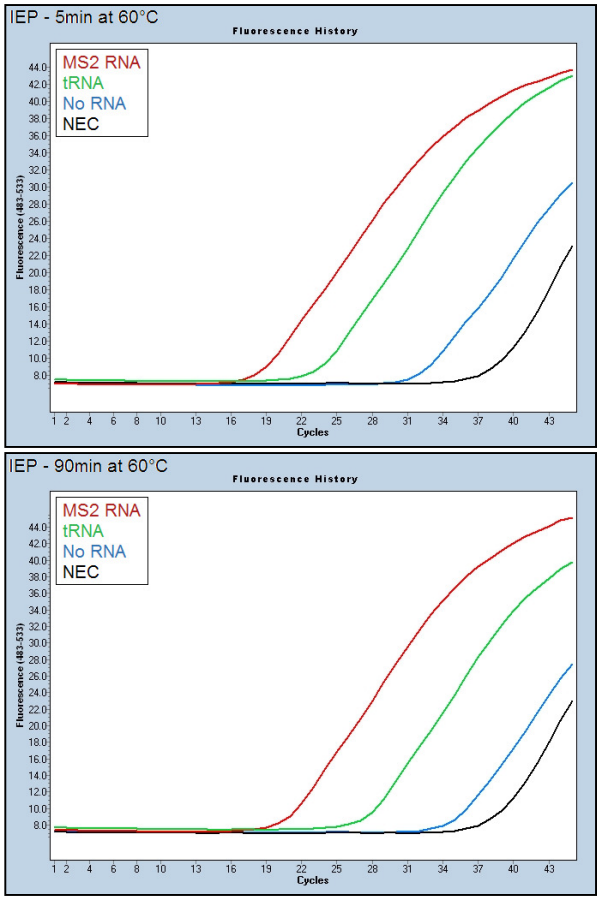

Figure 5.9: qPCR data showing the enzymatic activity of the IEP after 5min and 90min at 60ºC with and without the presence of RNA.

Incubation Time (min)	MS2 RNA (Cq value)	Yeast tRNA (Cq value)	No RNA (Cq value)	NEC (Cq value)
5	17	22	31	37
90	19	27	33	37

Table 5.4: A table showing the Cq values for an RT reaction using IEP after incubation of the enzyme for 5 and 90min at 60ºC in the presence of either MS2 RNA, yeast tRNA or no RNA.

The results showed that the RT reaction was stabilised by MS2 RNA with a Cq value of 17 cycles compared to that of 31 cycles when no RNA was present. The yeast tRNA also prevented reduction in enzymatic activity; however, the MS2 RNA seemed to have a greater stabilising effect on the RT. Since MS2 RNA stabilised the RT better than yeast tRNA and, due to the absence of primers during the incubation step, no cDNA synthesis could occur during the 60ºC incubation, the additional experiments were carried out in the presence of MS2 RNA.

MMLV-RT, IEP and IEP-Sac7d were all incubated in a cDNA synthesis reaction buffer, in the absence of primer, at 60ºC for varying time intervals. As before, primer was added and the first-strand reaction carried out at the RTs optimum reaction temperature for 30min followed by a 20min 95ºC heat denaturation step. The cDNA synthesis reaction was then analysed using qPCR (Figure 5.10 and Table 5.5).

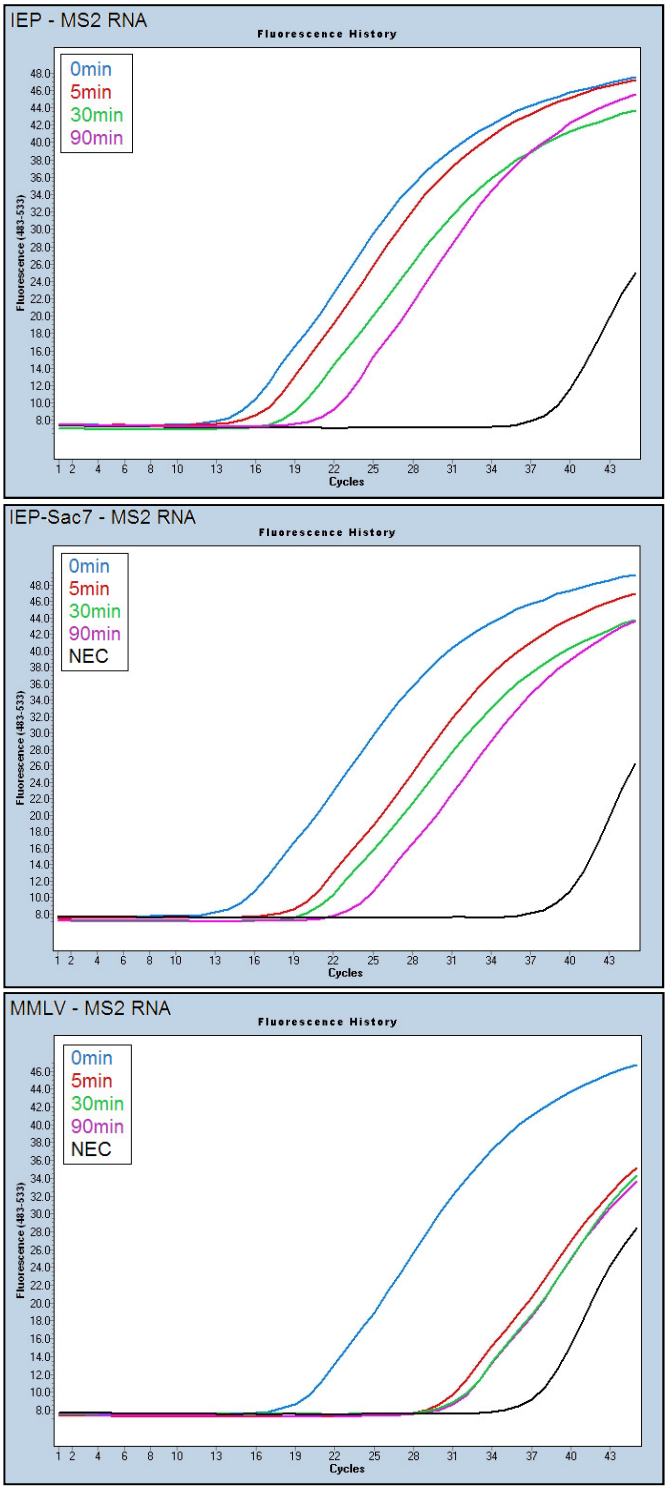

Figure 5.10: qPCR data showing the effect on enzyme activity of heating MMLV-RT, IEP and IEP-Sac7d for various times at 60ºC before being used to synthesise cDNA.

Incubation Time (min)	IEP (Cq value)	IEP-Sac7d (Cq value)	MMLV-RT (Cq value)
0	13	13	16
5	14	18	29
30	17	19	30
90	19	22	30
NEC	37	37	37

Table 5.5: A table showing the Cq values for RT reaction using IEP, IEP-Sac7d and MMLV-RT after incubation at set time intervals at 60ºC prior to the cDNA synthesis step.

The qPCR results for the 0min incubation differs from that seen in table 5.1 with lower Cq values being reported during this experiment. This is most likely due to a different set of primers being used, as primers with a high annealing temperature were selected for the earlier experiment, and due to a slightly different experimental conditions being carried out. Each reaction was set up at the same time before the 60ºC incubation occurred. It is possible that a slight loss of enzyme activity or RNA degradation was seen in the 0min control as the reaction was stored for 90mins before cDNA synthesis was allowed to occur. However, each enzyme in this experiment was treated the same and therefore any loss of activity seen would be consistent across the experiment.

The qPCR results show that the IEP's activity seems to be less affected by incubating at 60ºC than MMLV-RT. Using the IEP as the RT sees the Cq value increase by 1 cycle after 5min incubation; however, the Cq value for MMLV-RT under the same conditions increases by 13 cycles compared to no incubation. There is therefore a large reduction in the amount of cDNA being synthesised after the MMLV-RT has been incubated suggesting a loss of activity. The loss of activity with increasing incubation is progressive for the IEP, but even after 90min at 60ºC the Cq value for the IEP is at 19 cycles, lower than that seen for MMLV-RT after just 5min where the Cq is 29.

This experiment also shows that the Sac7d domain appears to be having a detrimental effect on the thermostability of the enzyme. This can be seen by analysis of the data after 5min incubation at 60ºC. The level of cDNA synthesised by the wild-type IEP was detected above the background level after 14 cycles; however, the addition of the Sac7d domain reduces the level of cDNA, increasing the Cq value to 18 cycles. The IEP-Sac7d is however still more thermostable than MMLV-RT and after 90min of incubation the amount of cDNA produced by the IEP is detectable after 22 cycles compared to 30 cycles for MMLV-RT. The most thermostable of the three RTs is the wild-type IEP, which produces cDNA after a 90min incubation that is detectable after 19 cycles

A control experiment was run in parallel with those mentioned here with the same temperature and time intervals but with no $MgSO_4$ present in the buffer. The relationships and trends mentioned above were the same, suggesting that the decline in cDNA synthesised with increasing incubation time was due to the loss of enzymatic activity and not Mg^{2+} catalysed hydrolysis of the RNA template. In addition, another experiment included 10µg BSA but no enhancement of enzymatic activity or reductions in the loss of activity were seen.

Fidelity Assay

To test the fidelity of the polymerases it was necessary to create two plasmids. The first plasmid, pCR®2.1_*lacI*^{q} would contain the *lacI*^{q} gene cloned into pCR®2.1. The T7 promoter site within pCR®2.1 would allow the production of *lacI*^{q} RNA to serve as a template for cDNA synthesis by the RTs. The second plasmid, pUC19_*sacB*, would contain the *B. subtilis sacB* gene, which would prove lethal when *E. coli* containing this construct was grown on LB containing 10%(w/v) sucrose.

Creating pCR®2.1 lacI^q Vector

The *lacI*^{q} gene was amplified using Phusion® DNA polymerase, pTTQ18NHK vector as template and primers designed to introduce *Sac* II sites at either end of the gene. Analysis on an agarose gel revealed a PCR product of approximately the correct size of 1298bp (Figure 5.11).

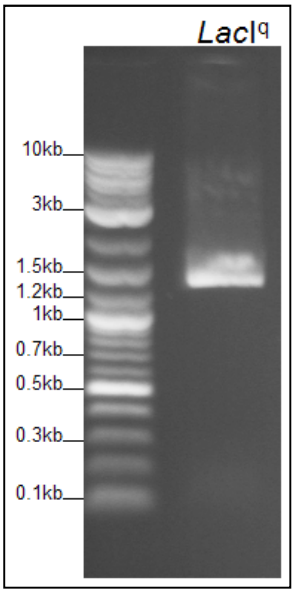

Figure 5.11: A 1%(w/v) agarose gel showing *LacI*^{q} gene amplified from pTTQ18NHK.

The amplified gene was purified and A-tailed to allow TA cloning into pCR®2.1. Blue-white screening on IPTG and X-Gal plates allowed the selection of positive colonies. These colonies were used to inoculate 5ml LB (with kanamycin) and grown overnight; the plasmid was then extracted using a Promega mini-prep kit. The plasmid DNA was sequenced using the ABI3100 protocol as it was essential that no errors were present in *lacI*^{q}.

Creating pUC19 sacB

sacB was amplified from pET24a_*sacB* using Phusion® DNA polymerase and primers designed to incorporate *Aat* II sites at either end of the gene. Analysis

on an agarose gel revealed a product approximately the correct size (1916bp) had been amplified (Figure 5.12).

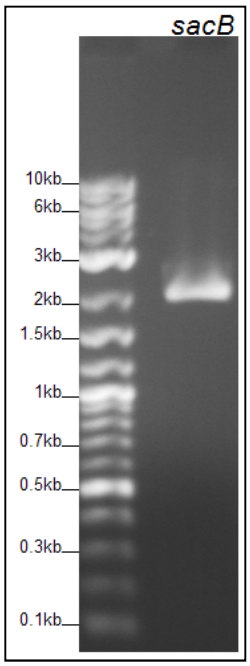

Figure 5.12: A 1%(w/v) agarose gel showing *SacB* amplified from pET24a_*sacB*

The purified *sacB* was digested with *Aat* II and ligated into pUC19 digested with the same enzyme. This construct was then transformed into TOP10 and plated onto LB agar plates containing methicillin and ampicillin. To ensure that sucrose lethality had been established, colonies were cross gridded onto plates that contained an additional 10%(w/v) sucrose. Colonies that would not grow on the sucrose plates were considered as positive for containing the *sacB* gene and were grown up on a 5ml scale; once their plasmids had been purified, the *sacB* gene was sequenced using the ABI3100 sequencing protocol to ensure no errors had been incorporated.

Proof of Principle

5ng of pCR®2.1_*lacI*^{q} was amplified with *Taq*, *Taq*:*Pfu* and Phusion® DNA polymerases using the *lacI*^{q}_*Sac*II primer pair. After 18 cycles, 5µl was

electrophoresed on an agarose gel with 0.5µg 2-log ladder to allow an estimation of the total yield (Figure 5.13).

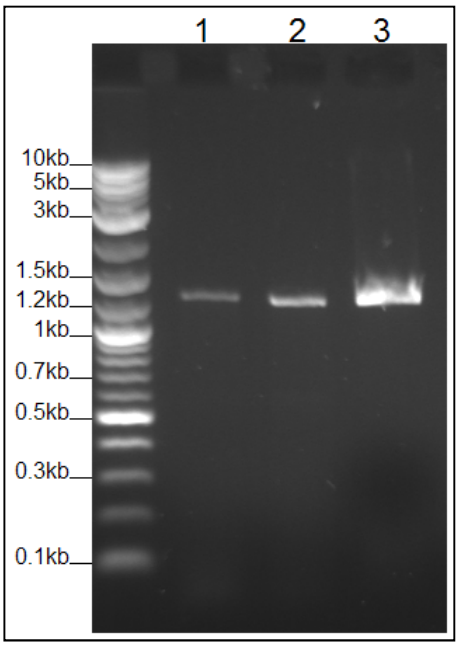

Figure 5.13: A 1%(w/v) agarose gel showing *Lac*I^{q} amplification with lane 1: *Taq* DNA polymerase, lane 2: *Taq*:*Pfu* (20:1) DNA polymerase mix and lane 3: Phusion® DNA polymerase

The purified *lac*I^{q} was digested with *Sac* II and ligated into pUC19_*sacB* digested with the same enzyme. This would allow the insertion of *lac*I^{q} into *sacB* interrupting the gene, thereby preventing lethality when grown on sucrose plates. This ensured that only constructs containing *lac*I^{q} would survive and therefore preventing a false background of blue colonies. The ligation constructs were transformed into electrocompetent TOP10 and plated on LB agar with antibiotic, X-Gal and sucrose. A control ligation containing just digested plasmid and no insert was also set up and transformed. No colonies appeared on the sucrose plates where no *lac*I^{q} insert was provided in the ligation, which confirmed the success of the sucrose lethality gene and allowed the blue:white colony ratio to be an accurate representation of *lac*I^{q} mutations.

Errors causing disruption of *lacI*^{q} would lead to a LacI^{-} phenotype and yield blue colonies on the X-Gal plates. The error rate was calculated with the equation $f = (-\ln F / d) \times 349$ (Frey and Suppmann, 1995) as detailed in the introduction to this chapter. Table 5.6 shows the results of this proof of principle experiment and, as expected, the *Taq* polymerase had the highest error rate at 2.66×10^{-4}, followed by the *Taq*:*Pfu* polymerase mix where the *Pfu* was providing some proof-reading and therefore reduced the error rate to 1.26×10^{-4}. Phusion® had the lowest error rate of 8.28×10^{-5} which was expected as Phusion® is marketed as a high-fidelity DNA polymerase. The error rates seen within this table are however higher than that reported for these enzymes. Finnzymes report Phusion to have a fidelity of 4.4×10^{-7} and *Taq* to have a fidelity of 2.28×10^{-5}. This discrepancy with error rate is quite large and possible due to a difference in the method used to measure fidelity. However, the general trend is as expected with the system capable of detecting differences between the DNA polymerases allowing a comparison of the fidelity of these enzymes.

Enzyme	Starting template (ng)	Yield (ng)	Duplications	White colonies	Total colonies	Fraction of white colonies	Calculated error rate (per bp)
Taq	5	250	5.64	4721	7967	0.5926	2.66×10^{-4}
Taq:*Pfu*	5	1000	7.64	6916	6988	0.7139	1.26×10^{-4}
Phusion®	5	4000	9.64	1261	1666	0.7569	8.28×10^{-5}

Table 5.6: The results of the proof of principle fidelity assay carried out on various DNA polymerases.

RT Fidelity

This system was adapted to be used to measure the fidelity of the various RTs studied in this report with potential problems being assessed and attempts made to reduce them. One problem incurred by this system stems from the fact that the T7 RNA polymerase could generate errors when generating the *lacI*^{q} RNA template; however, all the RTs would be using the same concentration of template from the same reaction and therefore should have an identical

baseline of errors. Additionally, the PCR step, after the cDNA synthesis, would introduce errors that could be magnified if more template was supplied from one RT than from another. It was therefore important to find the level of cDNA from each RT reaction by using a dilution series, to ensure the same concentration was amplified by the polymerase. It was also necessary to use a high-fidelity DNA polymerase, such as Phusion®, to minimise errors during this step.

The pCR®2.1_*lacI*^{q} vector was initially linearised with *Bgl* II before creating RNA using a T7 RNA polymerase. The reaction was treated with DNase I to ensure all template DNA had been removed and, once the DNase I had been heat-inactivated, a test PCR was carried out to confirm that no DNA remained in the RNA template sample. 50ng of the RNA was used in a first-strand reaction with AMV-RT, MMLV-RT, IEP and IEP-Sac7d for 30min at the optimum temperature, and then the reactions were heated to 95ºC for 20min to inactivate the RTs. PCRs were carried out on a dilution series of the cDNA template generated by each RT to ensure that the same amount of template would be amplified by the DNA polymerase, thereby reducing the effect of incorporated errors from the PCR. Once the same concentration had been established, the *lacI*^{q} was amplified from each cDNA template, purified and digested with *Sac* II. The digested fragment was then ligated into the *sacB* gene of pUC19*sacB*. Successful ligations would interrupt the *sacB* gene, allowing the colonies to survive when plated on LB containing 10%(w/v) sucrose. The constructs were transformed into electrocompetent TOP10 and plated on LB with antibiotic, sucrose and X-Gal and, after 24h at 37ºC, the number of blue and white colonies could be counted.

Since there are several steps during this experiment that, while they should be consistent in all the RTs tested, will add to the number of mutations, it was not possible to assign an accurate error rate for individual enzymes. However, the

percentage of white colonies could be calculated, allowing the RTs to be compared to each other. Table 5.7 shows the results of the fidelity assay.

Enzyme	Blue colonies	White colonies	Total colonies	White colonies (%)
AMV-RT	2490	4566	7056	64
MMLV-RT	947	1612	2559	65
IEP	2149	2825	4974	57
IEP-Sac7	3023	2913	5939	49

Table 5.7: Results of the fidelity assay when carried out on RTs using pCR®2.1_*lacI*^{q} as the RNA template.

The results show that the percentage of white colonies found in AMV-RT and MMLV-RT were very similar at 64% and 65%, respectively. However, the fidelity of the IEP and the IEP-Sac7d seems to be lower with 57% white colonies seen with IEP and 49% with IEP-Sac7d. The IEP is therefore not as accurate as the retroviral RTs with the Sac7d domain appearing to have an additional detrimental effect on the fidelity.

DNA-Dependent DNA Polymerase Activity

Retroviral RTs are known to show DNA-dependent DNA polymerase activity allowing them to convert their RNA genome into dsDNA. A FAM-labelled primer was used with a ssDNA template, M13mp8, and an extension event from the primer was measured using GeneScan software of an ABI3100 sequencer. A size standard GeneScan™ 350 ROX™ was used to allow accurate sizing of the extension products. As a positive control a known DNA-dependent DNA polymerase, *Taq*, was used to ensure that the system was working and that extension of the primer could be recorded. The result using 0.3125U of *Taq* was analysed using ABI Peak Scanner software where each peak represented a single extension product (Figure 5.14).

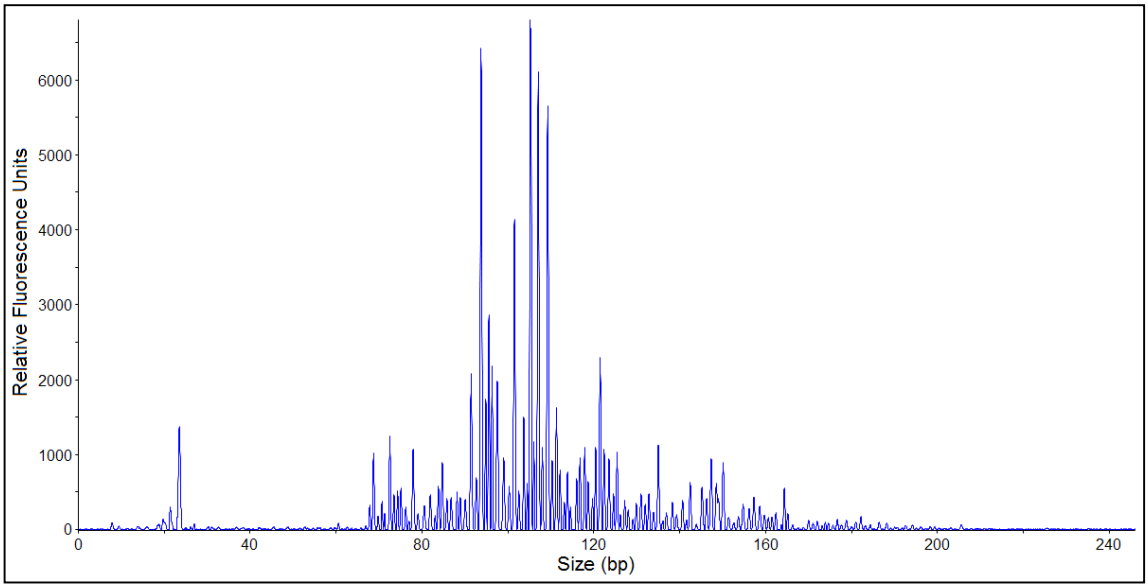

Figure 5.14: Processivity analysis of 0.3125U of *Taq* used to extend from a FAM-labelled primer annealed to M13mp8.

The initial part of the trace shows a peak at 23bp, which represents the FAM-labelled primer that has not been extended. When calculating the average size of the extension product it was important to omit this 23bp primer product from the final figure. From these results it was concluded that 0.3125U of *Taq*, one quarter of the amount that would normally be used in a PCR, extends on average 82 bases in 5min with a maximum extension event recorded of 179 bases.

It was important to run a negative-control with no enzyme to assess the base line that would be seen on a trace when no extension event occurs (Figure 5.15).

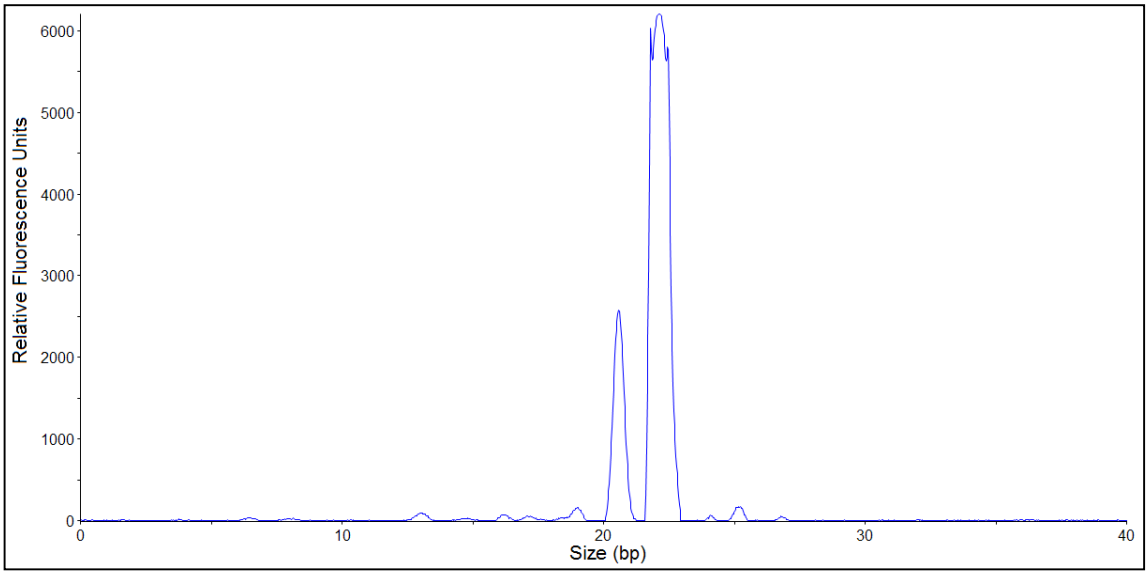

Figure 5.15: Processivity trace of the DNA-dependent DNA polymerase assay with no enzyme.

The major peak seen on the NEC represents the un-extended FAM-labelled primer and no significant extension products could be detected after this peak. The peak seen at 22bp is most likely due to the detection of a low level of incomplete primer synthesis with the loss of a base at the 3' end of the primer.

The experiment was repeated with AMV-RT. Since the retroviral RTs have low level DNA-dependent DNA polymerase activity, the reaction was carried out for 30min to ensure that a measurable extension event could occur (Figure 5.16).

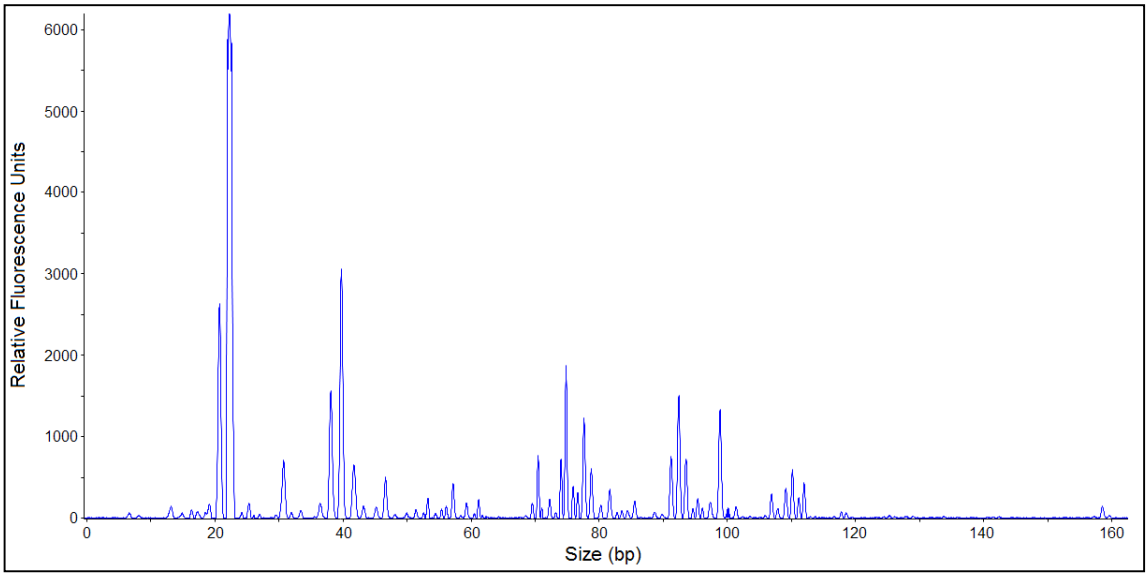

Figure 5.16: Processivity analysis of AMV-RT extending from a 23bp FAM-labelled primer for 30min using M13mp8 template.

The results of AMV-RT clearly show extension events after the 23 base primer peak. The 30min reaction showed that AMV-RT will extend, on average, 50 bases from the primer, with the maximum product detected at 135 bases.

The experiment was also repeated for MMLV-RT with a 30min extension time (Figure 5.17).

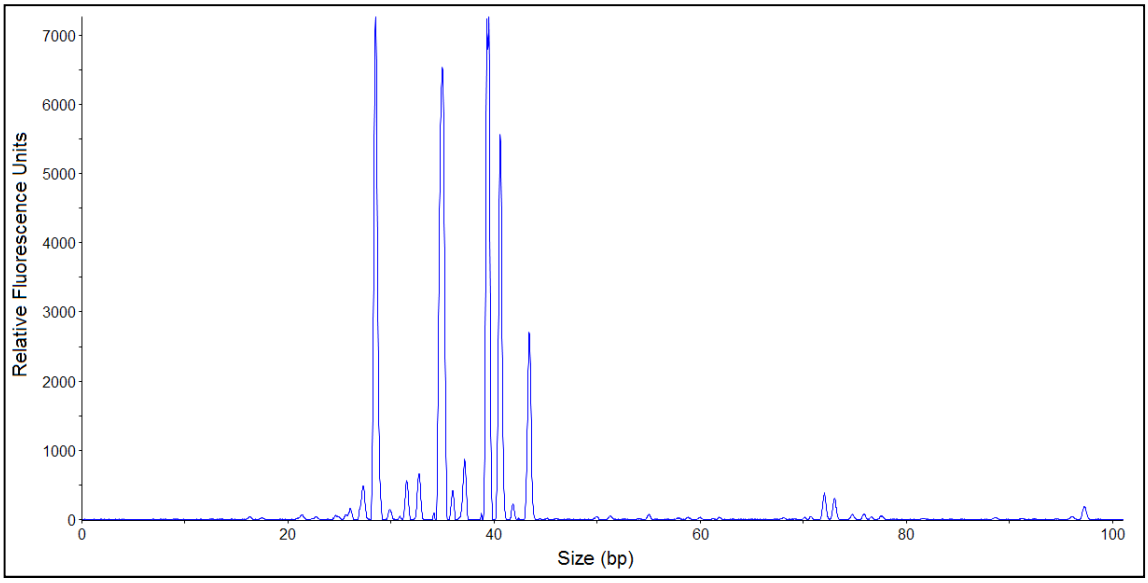

Figure 5.17: Processivity analysis of MMLV-RT extending from a 23bp FAM-labelled primer for 30min using M13mp8 template.

MMLV-RT showed less DNA-dependent DNA polymerase activity than AMV-RT and, although a maximum extension product was detected at 303 bases (not seen on the above graph), the average size product was 16 bases after a 30min extension time.

The experiment was repeated for both IEP and IEP-Sac7d with a 30min extension time (Figure 5.18 and Figure 5.19).

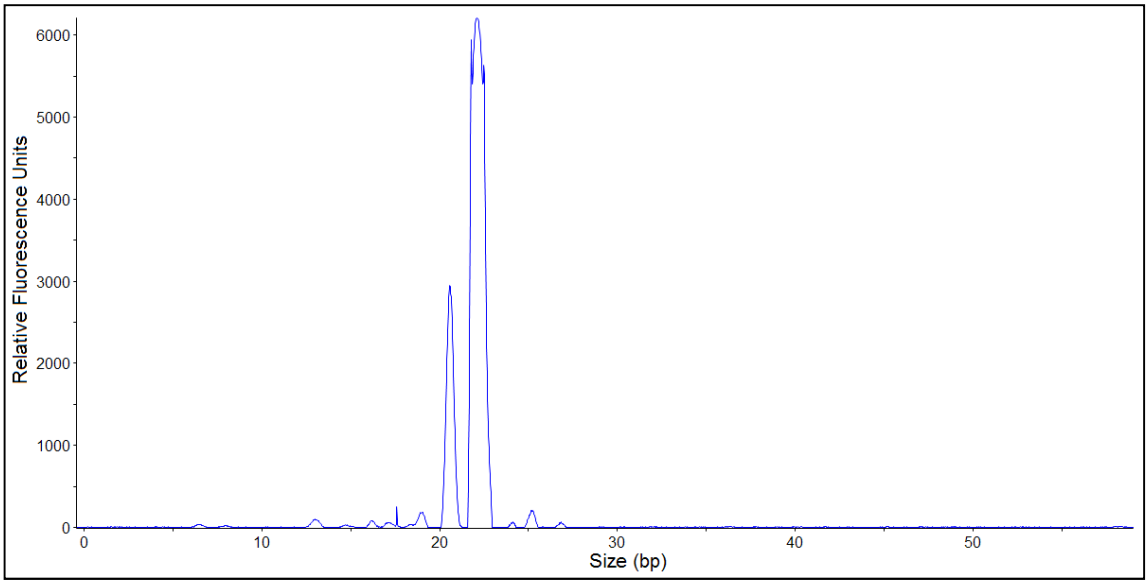

Figure 5.18: Processivity analysis of *B. caldovelox* IEP extending from a 23 base FAM-labelled primer for 30min using M13mp8 template.

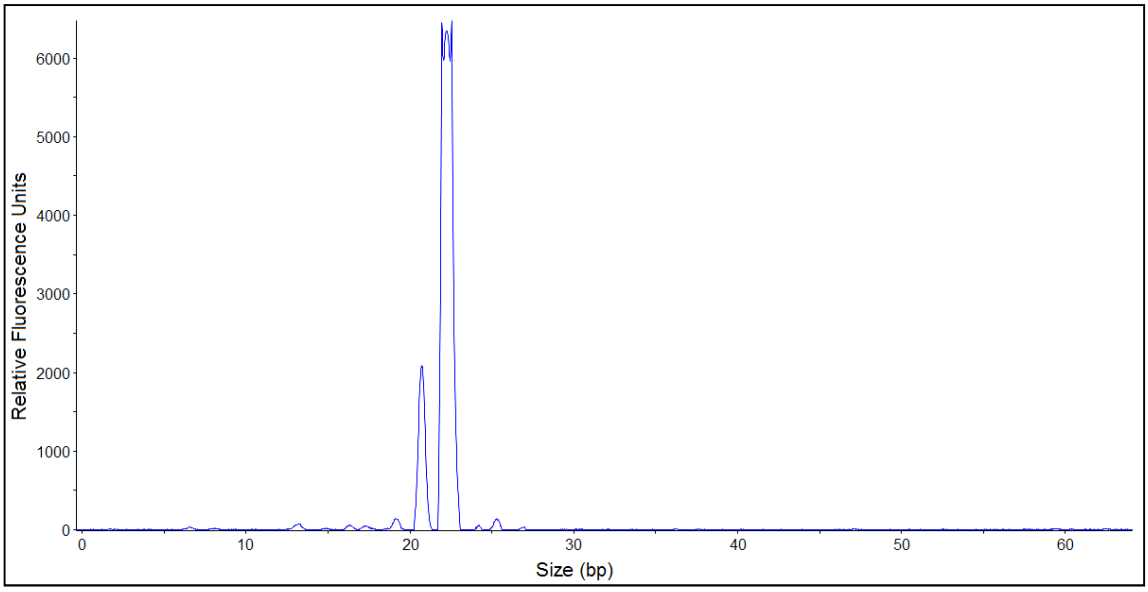

Figure 5.19: Processivity analysis of IEP-Sac7d extending from a 23 base FAM-labelled primer for 30min using M13mp8 template.

Both the IEP and IEP-Sac7d had profiles the same as that seen for the NEC and no significant peak could be detected after the 23 base primer. This

suggested that the IEPs, unlike retroviral RTs, do not exhibit DNA-dependent DNA polymerase activity.

RNA-Dependent DNA Polymerase Activity – Processivity

The DNA-dependent DNA polymerase assay was altered to allow the detection of cDNA extension products from an MS2 RNA template. The processivity can be calculated after repeated experiments altering both extension time and enzyme concentration until the median product length no longer changes (Wang *et al*. 2004). This allows the detection of a single enzymatic event and therefore would allow a comparison of processivity of IEP to retroviral RTs and to assess whether the Sac7d domain was improving the processivity of the IEP.

Initially, the experiment was set up using MMLV-RT as a control with an MS2 FAM-labelled primer, keeping all conditions the same as for the DNA-dependent DNA polymerase activity assay. However, no extension event could be detected on the trace. The experiment was repeated in several ways to include an experiment with:

- No primer annealing step in case the high 95ºC was degrading the RNA template.
- A primer-template annealing step with no magnesium present to reduce the likelihood of magnesium-catalysed RNA hydrolysis at the high initial primer annealing step

These attempts also failed to give any detectable extension product.

To ensure that the FAM-labelled primer was not interfering with the cDNA synthesis step, it was used in a standard first-strand reaction and then any cDNA produced was amplified using PCR. A correct sized product could be

detected both with the FAM-labelled primer and the same primer with no label suggesting that the primer was not inhibitory in any way. These PCR products were also electrophoresed on an agarose gel that was not stained with ethidium bromide and, using a dark reader, fluorescence was detected where the FAM-labelled primer was used, suggesting that this primer was correctly labelled so therefore should be detected by the sequencer.

Alternative attempts were made to include different purification methods with passing the reactions through a G-50 column instead of ethanol precipitation, but again no extension was detected. Additional attempts then included altering concentrations of the RNA template, both more dilute and more concentrated, different enzyme dilutions and altering incubation times. All failed to extend from the 24 base FAM-labelled primer.

No further attempts were made and this assay was not used to assess processivity of the RTs.

Basic processivity

Since the primer extension assay failed to produce processivity results, a more basic assay had to be developed. One feature of incorporating more bases in a single enzymatic event can be a faster enzyme due to the reduction in the occurrence of dissociation and reassociation of enzyme to the template. This was tested by incubating the IEPs at different time intervals and then using PCR to detect the length of product produced. Two different size products were amplified, a 1552bp (Figure 5.20) and 3407bp (Figure 5.21).

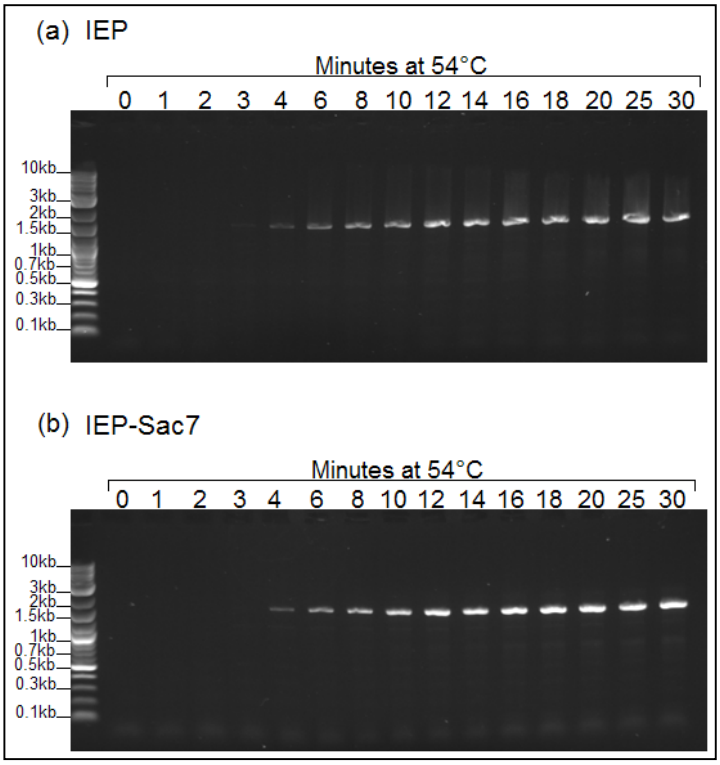

Figure 5.20: Results of the processivity assay of a) IEP and b) IEP-Sac7d when incubated with template for different time intervals and a 1552bp product amplified using PCR

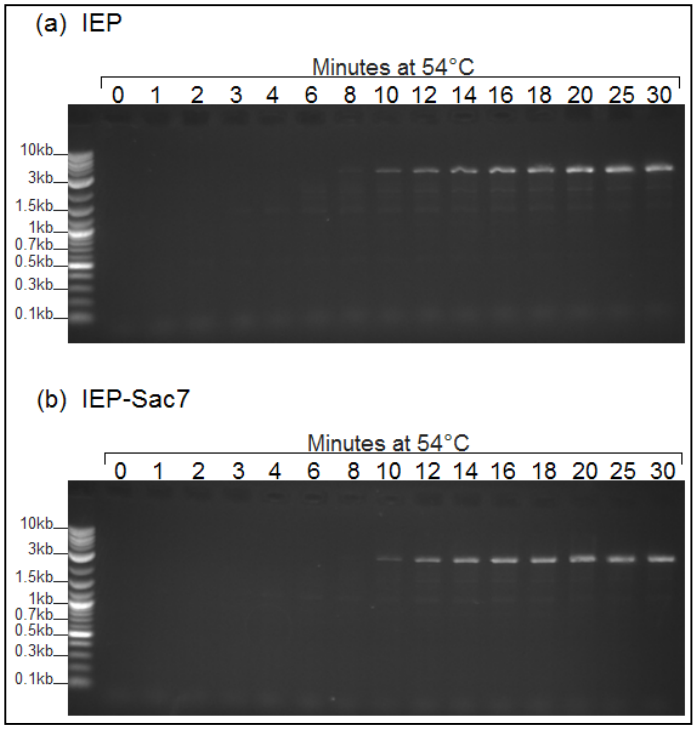

Figure 5.21: Results of the processivity assay of a) IEP and b) IEP-Sac7d when incubated with template for different time intervals and a 3407bp product amplified using PCR

Both enzymes required 4min to synthesise a target of 1552bp as detected by PCR and both required 10min to synthesise cDNA of 3407bp in length. The C-terminal Sac7d domain therefore does not seem to have any influence on the processivity of the enzyme during this assay as it is not synthesising the target in a reduced time. There is a possibility that the Sac7d is having a detrimental effect on the processivity of the enzyme as, when synthesising a 3407bp product, there is a very low level of full-length template detectable after 8min with the IEP that is not detectable with IEP-Sac7d. This basic processivity assay is therefore offering preliminary data suggesting that a C-terminal Sac7d domain will not enhance the enzyme.

Complex Target Assay

All the assays so far have been performed on basic MS2 RNA template, and therefore an additional assay was set up using total human placental RNA. A dilution series was set up of total human placental RNA and, using random nonamers, a cDNA synthesis step was carried out for 30min at 54ºC and 42ºC for IEP and MMLV-RT respectively. The cDNA template was then analysed in a qPCR. The qPCR results for 28S rRNA (Figure 5.22) showed that the IEP was not as efficient at producing cDNA as MMLV-RT. At 100ng of template the fluorescence became detectable above the background at 4 cycles with MMLV-RT. However, it is not until after 13 cycles that it becomes detectable for the IEP.

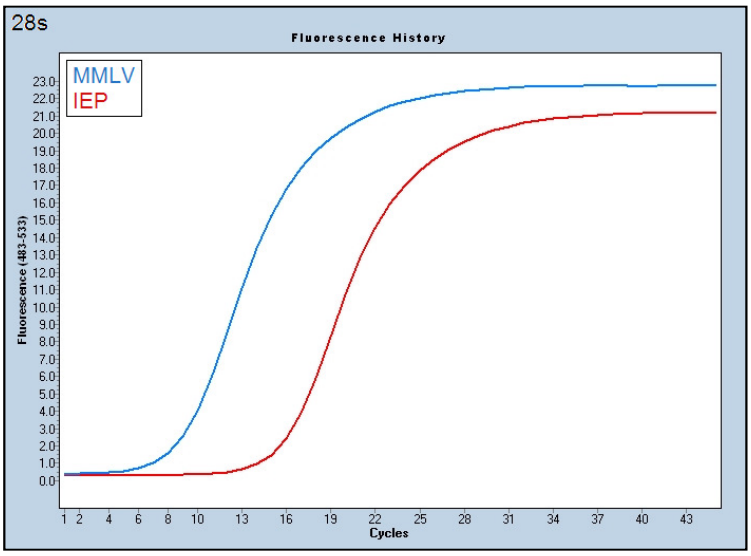

Figure 5.22: qPCR data comparing MMLV-RT (blue) to IEP (red) with 28S cDNA created from 100ng total human placental RNA.

The melt curve analysis showed that, although IEP is producing a measurable product that is detected much later, the product is indeed correct (Figure 5.23).

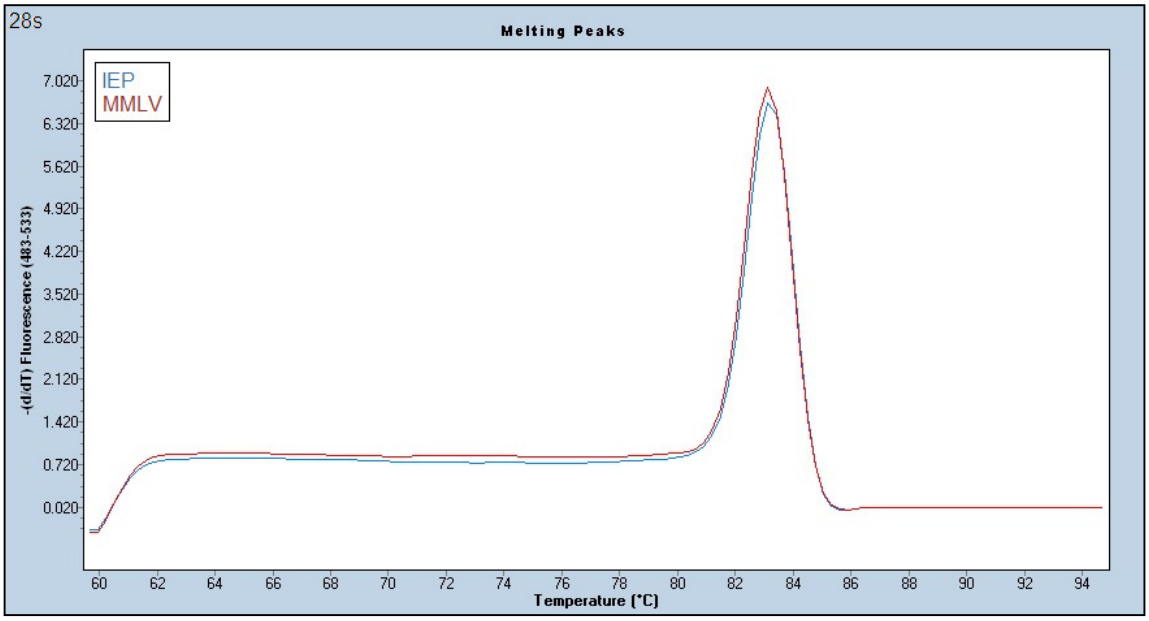

Figure 5.23: Melt curve analysis of the cDNA product created by both MMLV-RT and IEP with 100ng human RNA template and 28S primers.

Both the products from MMLV-RT and IEP melt at the same temperature of 83ºC suggesting that they are the same product despite the fact that more cycles are required to detect this product when an IEP is used as the RT.

The other three targets, GAPDH, ATP synthase and β-2-microglobulin mRNA all produce the same unusual results with IEP, from which GAPDH has been chosen as the example. GAPDH specific primers were used in a qPCR to amplify cDNA produced from a dilution series of RNA template using MMLV-RT, which represents what would be expected of a typical RT-qPCR. When compared to IEP it is clear that the amplification product with IEP is incorrect with no typical dilution series being seen (Figure 5.24).

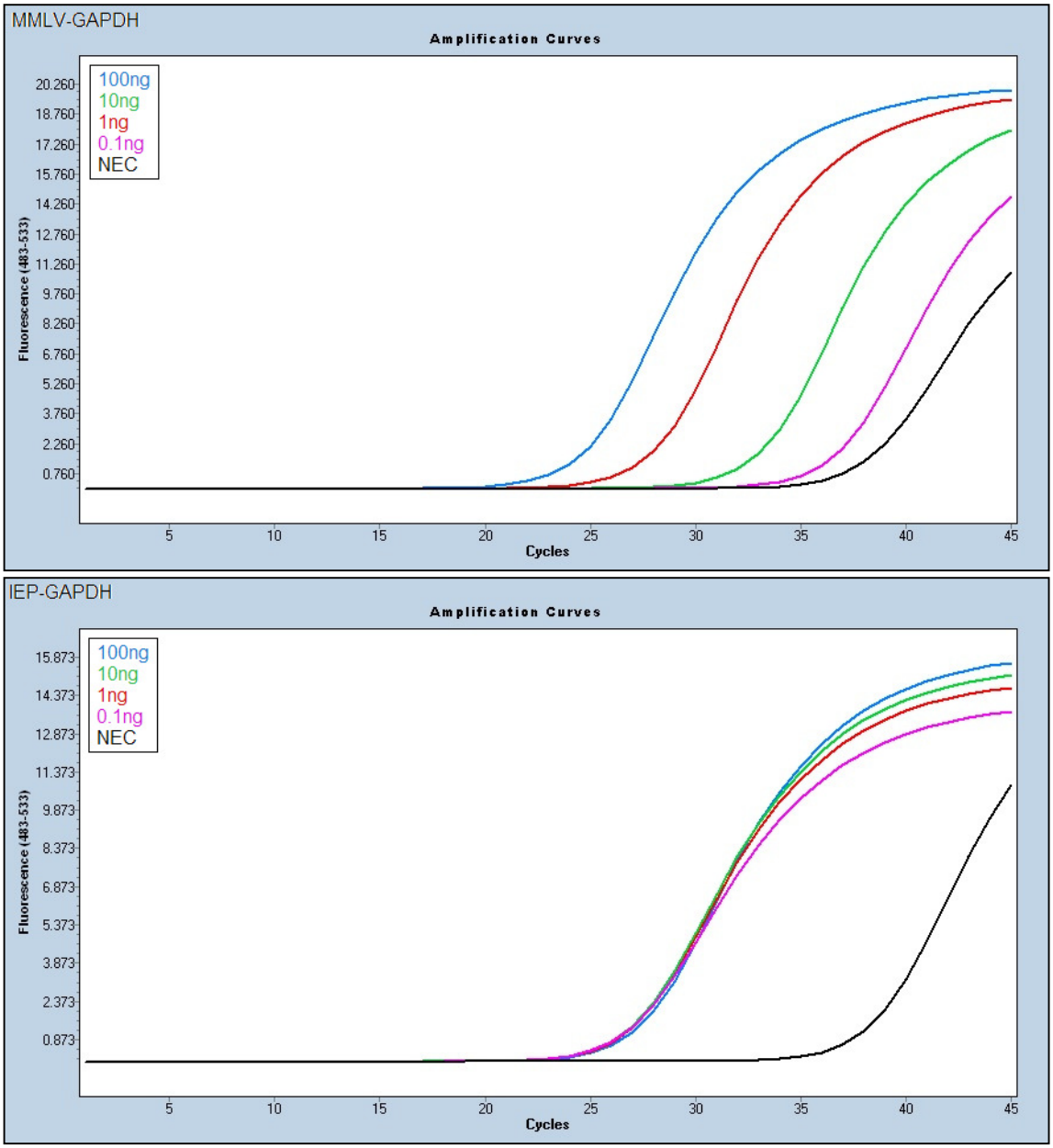

Figure 5.24: qPCR results of cDNA produced by MMLV-RT and IEP from an RNA dilution series with the qPCR primers used to target GAPDH.

The melt curve analysis of the qPCR reveal that IEP has produced an amplicon with a melting temperature of 76ºc compared to the correct product with an 82ºC melting temperature as seen with MMLV-RT (Figure 5.25). There is a low level of the correct product being synthesised by IEP as seen by the small peak at 82ºC; however, this was not the dominant amplicon of the reaction.

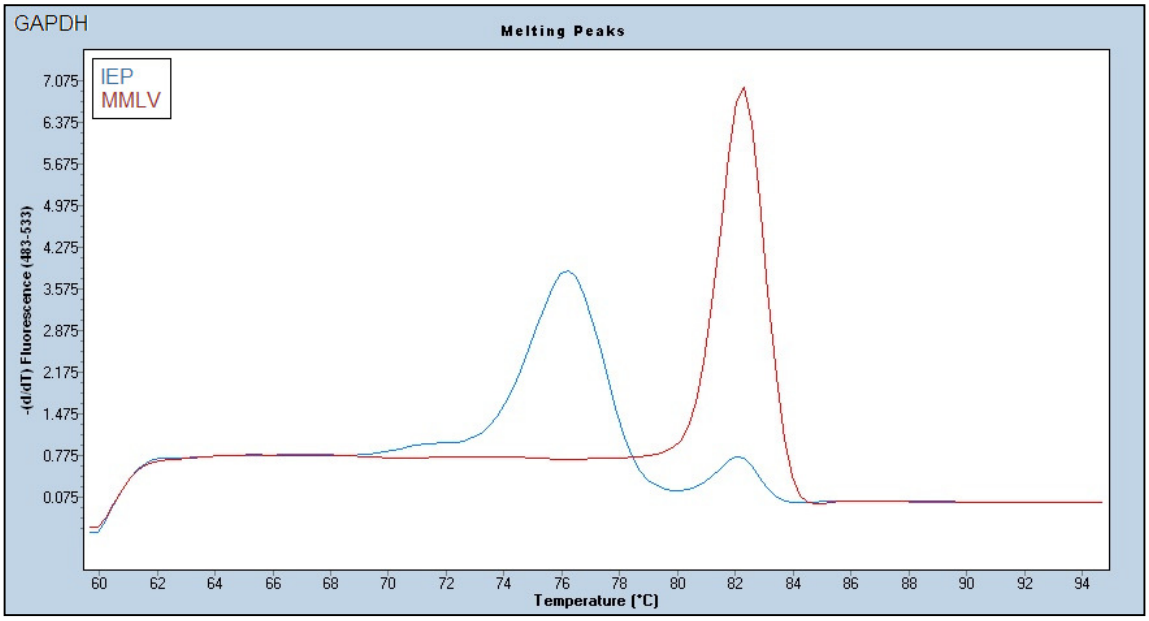

Figure 5.25: Melt curve analysis of the cDNA synthesis product from IEP and MMLV-RT using 100ng RNA and primers designed to target GAPDH

The experiment was repeated using random nonamers and Oligo (dT) primers with the reaction temperature reduced to 42ºC, eliminating the possibility of the primer not being able to anneal to the template. It was also repeated both at 42ºC and 54ºC using the specific primers in the cDNA synthesis step, but the same results were seen. This suggests that the IEP is unable to use complex targets such as total human placental RNA to produce cDNA unless the target is at a high-copy number as seen with 28S rRNA. The same results were seen with IEP-Sac7d, suggesting that the Sac7d domain is not enhancing the IEPs ability to use more complex targets.

5.4 – DISCUSSION

Characterisation of the IEP and IEP-Sac7d was carried out. Where necessary, the enzymes were compared to each other or to alternative commercially available retroviral RTs.

A nuclease assay showed that the purified IEP and IEP-Sac7d were free of any contamination that would interfere with the cDNA synthesis or subsequent PCR steps, and a dilution series of enzyme concentration revealed 1ng to be the optimum level of enzyme in a 20µl RT reaction.

Experiments to optimise the reaction buffers for the two IEPs revealed them to have no clear pH optimum with activity over the range tested. The main difference seen was that the Sac7d domain increased the IEPs requirement for salt with the optimum level of potassium sulphate increasing from 30mM for IEP to 80mM for IEP-Sac7d. This is a trend that has been observed before by GeneSys Ltd when using a Sac7d domain on other polymerases (unreported data) and it has been shown by Wang *et al.* (2004) that an Sso7d domain will actually increase a polymerase's tolerance to higher salt.

A cDNA step-temperature gradient allowed an optimum reaction temperature of 54ºC for both enzymes to be determined. This optimum was 12ºC higher than the optimum reaction temperature recommended for MMLV-RT. While the IEPs could function at higher temperatures, they showed more activity at 42ºC than at 64ºC. Although they did have a higher optimum reaction temperature than other retroviral RTs, this temperature was lower than expected as the bacterium *B. caldovelox* from which they were isolated has an optimum growth temperature

of 70ºC. It had therefore been expected that the RT would function at higher optimum temperatures than these data show. The IEPs however do have a higher optimum reaction temperature than current unmodified retroviral RTs.

When the IEPs were used to synthesise cDNA at temperatures higher than their optimum, they produced more cDNA than that seen with MMLV-RT. Although activity was still recorded at 76ºC for both IEPs and MMLV-RT, the IEPs were more active; the activity at 76ºC for the IEPs was higher than the recorded activity for MMLV-RT at 66ºC.

The IEPs were also tested to analyse what sources of divalent cation they could use for cDNA synthesis. Out of those tested, the IEP preferentially used sources of Mg^{2+} followed by $CoSO_4$ and then $MnCl_2$. Once the effect of the ions inhibiting the subsequent PCR was shown and reduced, additional experiments showed the IEPs to be able to use $NiCl_2$ and $CaCl_2$. At very low levels they could use $ZnSO_4$ but were unable to use $CuCl_2$ unless concentrations were specifically at 9mM. The IEPs are therefore broad in terms of divalent cation use.

To test for thermostability the enzymes were held under reaction conditions, but with no primer, for different time intervals at 60ºC. It was seen that in the absence of RNA the IEPS show a large loss of activity after only 5min of incubation. However RNA sources such as MS2 RNA or yeast tRNA improved the stability of the enzymes at higher temperatures, with MS2 RNA having a greater stabilising effect than tRNA. In the presence of MS2 RNA, the IEPs were more thermostable than MMLV-RT with more activity after 90min at 60ºC than MMLV-RT exhibited after just 5min. The IEPs were therefore acting as a

more thermostable RT enzyme capable of withstanding high temperatures for longer periods than MMLV-RT.

Initially, in order to test for fidelity, a proof of principle test on the DNA-dependent DNA polymerases was carried out. This showed that the method was sensitive enough to detect differences in the polymerases fidelity, showing that as expected *Taq* introduced the most errors, *Taq*:*Pfu* was slightly less error prone and that Phusion® DNA polymerase introduced the least errors.

The system was repeated for the retroviral RTs, AMV-RT and MMLV-RT, as well as for IEP and IEP-Sac7d. The error rate of the AMV-RT and MMLV-RT were very similar with 64% and 65% white colonies seen respectively. However, the error rate for the IEPs appeared to be higher with 57% and 49% white colonies seen. A low fidelity rate is expected for the retroviral RTs since mutations can be advantageous in allowing the virus to evolve and adapt, as seen with the low fidelity of HIV-1 RT (Ji and Loeb. 1992). However, Ng *et al*. (2007) expected the IEPs to have some level of fidelity to reduce the likelihood of the intron being unable to splice from the RNA transcript and therefore prove deleterious to the cell.

The RTs were also tested for DNA-dependent DNA polymerase activity. It is known that retroviral RTs do have this activity, enabling them to produce dsDNA from their RNA genome (as reviewed by Herschhorn and Hizi, 2010). Our results showed detectable levels of this activity using retroviral RTs, with AMV-RT showing more activity than MMLV-RT. Testing the IEPs, however, revealed them not to exhibit any detectable DNA-dependent DNA polymerase activity even after 30min incubation time. This is possibly due to the fact that the intron

can utilise the host cell DNA polymerase to allow its re-insertion back into the genome and therefore does not require this activity.

In order to analyse the influence of a Sac7d domain on an IEP, an assay was attempted to measure RNA-dependent DNA polymerase activity, specifically the number of nucleotides incorporated in a single enzymatic event. Despite the fact that this method worked on both DNA polymerases and RTs for DNA-dependent DNA polymerase activity and the fact that all components were tested for inhibitory effect, this assay did not produce detectable extension events. Several variables were altered including template preparation, template purification, enzyme concentration, incubation times and template concentrations. However, it was not possible to measure processivity using this assay. A more basic assay that measured the speed of the IEP revealed no significant difference when a Sac7d domain was present on the IEP.

The characterisation of the *B. caldovelox* IEP has revealed the enzyme to be more thermostable than unmodified MMLV-RT and that, using basic targets, is more efficient at temperatures above its optimum than MMLV-RT. However, the enzyme is unable to synthesise cDNA efficiently when more complex templates are used, as seen with assays to amplify specific regions from total human placenta RNA. Interestingly, unlike retroviral RTs, this IEP shows no DNA-dependent DNA polymerase activity and due to a lack of RNase H domain presumably shows no RNase H, activity reflecting the difference in biological function of the IEP compared to the retroviral RTs.

Chapter 6 – Summary, Discussion and Future Aspects

6.1 – REPORT SUMMARY

This Ph.D report introduces the problems that can be incurred when using currently available RTs for research and diagnostic purposes. In order to reduce these problems it was decided that an RT with a high optimum reaction temperature would be required. A search for thermophilic RTs was carried out from both un-sequenced environmental bacteria and by carrying out BLAST searches against available genome sequences from thermophiles. Using this method several previously uncharacterised RTs, from IEP within Group II Introns, were discovered. These included IEPs from *B. caldovelox* and *B. caldotenax*, which were identical in gene sequence, one from a New Zealand strain of *B. stearothermophilus*, one from *T. carboxydivorans* and three almost identical IEPs from *P. mobilis*.

The cloning of these genes followed by their protein expression was successfully carried out and a manipulation of the *B. caldovelox* IEP was made with the fusion of a Sac7d domain on the C-terminal of the protein. Purification of these proteins proved to be very difficult as the enzyme co-purified with many protein contaminants or the removal of nucleic acids causing protein insolubility. Despite using an extensive range of methods to purify the IEP from *T. carboxydivorans* and one of the IEPs from *P. mobilis*, a pure protein sample was never obtained. However, both partially purified enzymes were shown to exhibit RT activity using MS2 RNA as a template.

The IEP from *B. caldovelox* and the IEP-Sac7d fusion protein were successfully purified and characterisation studies were carried out. Both enzymes were shown to have an optimum reaction temperature at 54ºC. This optimum temperature was 12ºC higher than the recommended optimum reaction temperature of MMLV-RT. The RTs could also synthesise cDNA at temperatures higher than their optimum with greater efficiency than MMLV-RT, with activity still being seen at 76ºC. This higher reaction temperature is potentially very useful in situations where mesophilic RTs have struggled to synthesise cDNA from targets with a high degree of secondary structure.

Retroviral RTs also display DNA-dependent DNA polymerase and RNase H activity, which are required as part of their retroviral lifecycle. The lack of an RNase H domain on the IEPs suggests that no RNA degradation would occur within an RNA:DNA hybrid. This is an advantage that would prevent cleavage of the RNA template and therefore reduce truncated products in the final cDNA yield. The DNA-dependent DNA polymerase assay revealed a complete absence of this activity within the IEPs, reflecting the difference in their lifecycle compared to the retroviral RTs.

The wild-type *B. caldovelox* IEP was compared to IEP-Sac7d to see if there was any enhancement in the processivity of the enzyme with the addition of the Sac7d domain. However, the sensitive assay used to detect DNA-dependent DNA polymerase activity was unsuccessful when using RNA, with several different approaches failing to detect an extension product. A more basic processivity assay showed that the Sac7d domain appeared to have no enhancement on the speed of the enzyme. As a further disadvantage, this additional domain also appeared to lower the thermostability of the enzyme with a slight reduction in activity seen on qPCR data with increasing reaction temperatures.

Although the *B. caldovelox* IEP has potential as a thermophilic RT enzyme a major problem was seen when challenging the enzymes with low-copy targets and complex templates in the form of human placental RNA. The IEPs were incapable of producing correct cDNA under several different conditions where MMLV-RT was successful.

6.2 – DISCUSSION AND FUTURE ASPECTS

Optimum Reaction Temperature

The studied IEP of *B. caldovelox* had a higher optimum reaction temperature than wild-type retroviral RTs. However, at 54ºC this optimum reaction temperature was at least 16ºC lower than expected. The growth temperature of the bacterium is 70ºC, and as a general rule enzymes are usually adapted to be functioning optimally in a range close to or slightly above that of the growth conditions of the source organism (Hough and Danson, 1999). The IEPs from *B. caldovelox, B. caldotenax* and *G. kaustophilus* were all identical and it is possible that, due to the mobile nature of Group II Introns, these IEPs were originally from a bacterium that grew at lower temperatures. These thermophilic bacteria could have acquired the Group II Intron, containing the IEP, from this mesophilic bacterium. Even if the IEP was not functional at their growth temperatures, the Group II Intron would not prove deleterious to the host providing that the intron was still capable of auto-splicing from RNA transcripts. With no known advantage to the host, there would be no pressure on the IEP to adapt to the higher temperature and it therefore would not necessarily be a thermostable enzyme.

Improving *B. caldovelox* IEP Performance

The IEP from *B. caldovelox* looks disappointing in terms of its potential for commercial use as an RT. Chapter 5 highlighted the fact that this IEP has difficulties using complex RNA templates. More research needs to be carried out on the IEP to detect the reason for the poor performance on these templates. Several factors could be influencing the IEP during these experiments, including:

- The IEP may not be sensitive enough to detect low-copy RNAs or these low-copy RNAs are destroyed during the higher reaction temperature

- There is a possibility that a low level of *E. coli* RNA is still present in the sample. This low level of RNA could be providing a non-specific target during cDNA synthesis, which is why a standard template dilution series cannot be detected. Running a no-template control with the enzyme and using *E. coli* specific primers could indicate if this was the case.
- Although 1ng of enzyme is required for the basic MS2 RNA template, a much larger quantity might be optimum for a more difficult and complex target. Altering reaction component concentrations may see an improvement in the performance of this enzyme.

Fidelity

Unlike retroviruses, the Group II Introns do not need to evade the immune system of a eukaryotic host and therefore do not have this pressure to constantly mutate and evolve. It was expected that the Group II Intron would require a high-fidelity RT to reduce the likelihood of a mutation occurring that could prevent the auto-splicing of the intron from an RNA transcript. If the intron permanently interrupted an essential host gene then this could prove deleterious or possibly lethal to the host. It was a surprise to find that the error rate of the *B. caldovelox* IEPs were in fact lower than that seen for both AMV-RT and MMLV-RT (chapter 5). This lower fidelity would be problematic where an exact copy of the mRNA transcript is required for research such as sequencing or molecular cloning, and therefore poses a huge disadvantage to the use of these RTs in the place of current retroviral RTs.

Characterisation of Alternative IEP

More promising thermophilic RTs could still be found in the IEPs of *T. carboxydivorans* and *P. mobilis*. Although these partially purified proteins were shown to exhibit RT activity, they were not analysed in terms of their optimum reaction temperature or thermostability due to the impurity of the sample. It is

possible that these enzymes could have a higher optimum reaction temperature than the IEPs studied so far, and therefore the complete purification of these proteins is essential to continue the research on these types of RTs.

Alternative Purification Methods

A major limitation in the purification of some of the IEPs appears to be related to their tight association with nucleic acids. An experiment carried out during the purification process showed that the *T. carboxydivorans* protein was associating specifically with RNA and not DNA (Chapter 4). This can be expected as with an absence of DNA-dependent DNA polymerase activity the protein has no requirement for binding DNA. A method to efficiently remove this RNA contamination is required, possibly by using RNase immobilised on a column. However, it would be essential that the column was in no way 'leaky' contaminating the protein sample with RNases that would inhibit subsequent reactions by cleaving the RNA template.

Completely removing nucleic acids from the sample might prove impossible. It could be that nucleic acids are essential for the stability of these enzymes. As seen in Chapter 5, the absence of RNA during a 60ºC incubation step resulted in a marked decrease in enzymatic activity compared to the same experiment in the presence of RNA. The RNA is clearly having an effect on the enzyme, causing it to be more stable at higher temperatures, and could be required for the other IEPs that were attempted to be purified to ensure they remain soluble. In chapter 4 it was found that in partially pure samples of *T. carboxydivorans* IEP, in the absence of nucleic acids, insoluble material formed and RT activity could no longer be detected. This would be a great limitation to the purification of these enzymes as it is essential that background nucleic acids are removed to avoid contamination of downstream applications.

Processivity Studies

Processivity was also explored during this report. Having a more processive enzyme could reduce reaction times at the high temperatures, therefore minimising metal ion catalysed hydrolysis of the mRNA template (Das *et al*. 2001). Furthermore, a tighter association with the template by means of attaching a tethering protein to the template could reduce complete dissociation of the enzyme, so reducing the level of truncated products in the final yield. Sso7d is a DNA-binding protein that has been shown to improve processivity of DNA-dependent DNA polymerases, presumably by tethering the protein to the DNA strand and reducing the occurrence of dissociation events (Wang *et al*. 2004). Preliminary results on the Sac7d domain fused to the IEP showed no apparent enhancement of the enzyme in terms of its processivity. In Nature the Sac7d domain would bind dsDNA and is possibly not able to have the same tethering affect, seen with DNA polymerases, due to the RNA template and the RNA:DNA hybrid constructs.

It is possible that a less specific molecule would be needed to improve the processivity of these types of polymerases. Thioredoxin, for example is an alternative protein that is thought to act as a clamp once it has bound the polymerase (Bedford *et al*. 1997). This protein has no natural affinity for DNA and this lack of specificity could allow the method to be extended to RTs, enabling the enzyme to produce longer length cDNA during a single enzymatic event.

Another alternative enhancement factor for RTs might be found in the form of bacterial ssDNA-binding proteins (SSB) instead of the dsDNA-binding protein, Sac7d, which has been studied in this report. Their role in Nature is thought to be to protect DNA where the double-stranded structure has been disrupted (Perales *et al*. 2003). The *Thermus thermophilus* SSB (*Tth*SSB) was studied by Perales *et al*. (2003) as a possible method to enhance PCR. This group found

that the *Tth*SSB protein not only halved elongation time for *Tth* and *Pfu* DNA polymerases, but that it also had an additional effect of increasing the fidelity of *Tth* DNA polymerase. Their results then went on to show that the *Tth*SSB protein could also interact with RNA and greatly enhanced the RT activity of the *Tth* DNA polymerase allowing the production of cDNA, detected by PCR, up to 3285 bases in length, which could not be achieved in the absence of *Tth*SSB. SSBs therefore have a potential role in improving processivity of RT-PCR when added during the cDNA synthesis step. The affect of enhanced fidelity on RT activity is yet to be studied.

Effect of Template Related RT Pausing

The IEP from *B. caldovelox* has yet to be tested for dissociation at sites of secondary structure, partly due to the failure of the processivity assay. Once this assay is functional it could be adapted to use a FAM-labelled primer and an RNA template with known secondary structure to detect whether dissociation events are common at this site. Frequently occurring dissociation events at a specific site will result in the accumulation of known sized products that can be detected and recorded as a major peak using the ABI3100 Gene Scanner programme and Peak Scanner Software. Comparisons can be made with retroviral RTs and with increasing reaction temperatures to see if a thermostable enzyme will allow the reduction or elimination of these dissociation events.

Alternative Characterisation

Currently there is limited information as to the exact structural nature of the IEP from a Group II Intron. Blocker *et al*. (2005) compared a theoretical secondary structure of the LtrA protein encoded by the *Lactobacillus lactis* LI.LtrB Group II Intron and predicated the protein to form a homodimer. However, even among related RTs such as those from retroviruses the structures can differ, with MMLV-RT forming a monomer and AMV-RT forming a heterodimer.

Techniques such as gel filtration will allow an estimation of the Mr value of the protein and therefore elucidation of the basic protein conformation by means of its oligomeric state while X-ray crystallographic studies could provide more detail into the protein structure as well as protein-template interactions and possibly structural roles of the additional protein domains found at the C-terminus of the protein.

REFERENCES

- Baltimore, D. 1970. Viral RNA-dependent DNA polymerase. *Nature* 226, 1209-1211
- Bedford, E., Tabor, S., Richardson, C.C. 1997. The thioredoxin binding domain of bacteriophage T7 DNA polymerase confers processivity on *Escherichia coli* DNA polymerase I. *Proc. Natl. Acad. Sci. USA.* 94, 479-484.
- Belancio, V.P., Hedges, D.J., Deininger, P. 2008. Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. *Genome Research* 18, 343-358
- Berger, S.L., Wallace, D.M., Puskas, R.S., Eschenfeldt, W.H. 1983. Reverse Transcriptase and Its Associated Ribonuclease H: Interplay of Two Enzyme Activities Controls the Yield of Single-Stranded Complementary Deoxyribonucleic Acid. *Biochemistry* 22, 2365-2372.
- Blocker, F.J.H., Mohr, G., Conlan, L.H., Qi, L., Belfort, M., Lambowitz, A.M. 2005. Domain structure and three-dimensional model of a Group II Intron-encoded reverse transcriptase. *RNA* 11, 14-28.
- Bramucci, M.G., Nagarajan, V. 1996. Direct Selection of Cloned DNA in *Bacillus subtilis* Based on Sucrose-Induced Lethality. *Applied and Environmental Microbiology* 62, 3948-3953.
- Brown, T.A. 2001. *Gene Cloning and DNA Analysis*. 4th edition. Oxford: Blackwell Science Ltd.
- Buell, G.N., Wickens, M.P., Payvar, F., Schimke, R.T. 1978. Synthesis of full-length cDNAs from four partially purified oviduct mRNAs. *The Journal of Biological Chemistry* 253, 2471-2482
- Bustin, S.A. 2004. *A-Z of Quantitative PCR*. California: IUL Biotechnology Series.
- Castro, C., Smidansky, E., Maksimchuk, K.R., Arnold, J.J., Korneeva, V.S., Götte, M., Konigsberg, W., Cameron, C.E. 2007. Two proton

transfers in the transition state for nucleotidyl transfer catalyzed by RNA- and DNA-dependent RNA and DNA polymerases. *Proceedings of the National Academy of Science, USA* 104, 4267-4272.

- Collins, R.A., Stohl, L.L., Cole, M.D., Lambowitz, A.M. 1981. Characterisation of a novel plasmid DNA found in mitochondria of *N. Crassa*. *Cell*. 24, 443-452
- Crick, F. 1970. Central Dogma of Molecular Biology. *Nature* 227, 561-563.
- Das, M., Harvey, I., Cuh, L.L., Sinha, M., Pelletier, J. 2001. Full-length cDNAs: More than just Reaching the Ends. *Physiol Genomics*. 6, 57-80
- Darlix, J., Lapadat-Tapolsky, M., Rocquigny, H., Roques, B.P. 1995. First Glimpses at structure-function Relationships of the Nucleocapsid Protein of Retroviruses. *Journal of Molecular Biology* 254, 523-537
- Davidson, J.F., Fox, R., Harris, D.D., Lyons-Abbott, S., Loeb, L.A. 2003. Insertion of the T3 DNA polymerase thioredoxin binding domain enhances the processivity and fidelity of *Taq* DNA polymerase. *Nucleic Acids Research* 31, 4702-4709.
- DeStafano, J.J., Buiser, R.G., Mallabert, L.M., Myers, T.W., Bambara, R.A., Fay, P.J. 1991. Polymerization and RNase H activities of the reverse transcriptases from Avian Myeloblastosis, Human Immunodeficiency, and Moloney Murine Leukaemia Viruses are functionally uncoupled. *The Journal of Biological Chemistry* 266, 7423-7431.
- DSMZ online: http://www.dsmz.de/microorganisms/html/strains/strain.dsm000411.html [accessed February 21 2011]
- EcoProDB. 2006. Theoretical Map. KAIST Institute for the BioCentury. Available from http://eecoli.kaist.ac.kr/theoretical.html [Accessed 3rd November 2010]
- Eickbush, T.H. 1997. Telomerase and Retrotransposons: Which came first? *Science* 277, 911-912

- El-Diery, W.S., Downey, K.M., So, A.G. 1984. Molecular Mechanisms of Manganese Mutagenesis. *Proc. Natl. Acad. Sci. USA.* 81, 7378-7378.
- Fedorova, O., Zingler, N. 2007. Group II Introns: Structure, Folding and Splicing Mechanisms. *Biol Chem*. 388, 665-678.
- Flavell, A.J. 1995. Retroelements, reverse transcriptase and evolution. *Comp. Biochem. Physiol.* 110, 3-15.
- Filippo, J.S., Lambowitz, A.M. 2002. Characterisation of the C-terminal DNA-binding/DNA endonuclease region of a Group II Intron-encoded protein. *Journal of Molecular Biology* 324, 933-951.
- Freeman, W.M., Vrana, S.L., Vrana, K.E. 1996. The use of Elevated Reverse Transcription Reaction Temperatures in RT-PCR. *BioTechniques* 20, 782-783.
- Filler, A.G., Lever, A.M.L. 1997. Effect of Cation Substitutions of Reverse Transcriptase and on Human Immunodeficiency Virus Production. *AIDS Research and Human Retroviruses.* 13, 291-299
- Frey, B., Suppmann, B. 1995. Demonstration of the Expand™ PCR System's Greater Fidelity and Higher Yields with a *lac*I-based PCR Fidelity Assay. *Biochemica* Technical Tips. 2, 8-9
- Fuchs, B., Zhang, K., Rock, M.G., Bolander, M.E., Sarkar, G. 1999. High Temperature cDNA synthesis by AMV Reverse Transcriptase Improves the Specificity of PCR. *Molecular Biotechnology.* 12, 237-240
- Fusi, P., Tedeschi, G., Aliverti, A., Ronchi, S., Tortora, P., Guerritore, A. 1993. Ribonuclease from the extreme thermophilic archaebacterium *S. Solfataricus*. *FEBS*. 211, 305-310
- Gay, P., Le Coq, D., Steinmetz, M., Berkelman, T., Kado, C.I. 1985. Positive Selection Procedure For Entrapment of Insertion Sequence Elements in Gram-Negative Bacteria. *Journal of Bacteriology*. 164, 918-921
- Gerard, G.F., Fox, D.K., Nathan, M., D'Alessio, J.M. 1997. Reverse Transcriptase: The use of cloned Moloney murine leukaemia virus

reverse transcriptase to synthesize DNA from RNA. *Molecular Biotechnology* 8, 61-67

- Goodman, M.F., Keener, S., Guidotti, S. 1983. On the Enzymatic Basis for Mutagenesis by Manganese. *The Journal of Biological Chemistry* 258, 3469-3475.
- Götz, D., Banta, A., Beveridge, T.J., Rushdi, A.I., Simoneit, B.R.T., Reysenbach, A.L. 2002. *Persephonella marina* gen. nov., sp. nov. and *Persephonella guaymasensis* sp. nov., two novel, thermophilic, hydrogen-oxidizing microaerophiles from deep-sea hydrothermal vents. *International Journal of Systematic and Evolutionary Microbiology* 52, 1349-1359.
- Gogvadze, E., Buzdin, A. 2009. Retroelements and their impact on genome evolution and functioning. *Cell. Mol. Life Sci.* 66, 3727-3742
- Gravelat, F.N., Doedt, T., Chiang, L.Y., Liu, H., Filler, S.G., Patterson, T.F., Sheppard, D.C. 2008. In Vivo analysis of *Aspergilus fumigates* Developmental Gene Expression Determined by Real-Time Reverse Transcription-PCR. *Infection and Immunity* 76, 3632-3639
- Griffiths, A.J.F. 1995. Natural Plasmids of Filamentous Fungi. *Microbiological Reviews* 59, 673-685
- Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R. 1999. Correlation between protein and mRNA abundance in Yeast. *American Society for Microbiology* 19, 1720-1730
- Haig, H., Kazazian, Jr. 2004. Mobile Elements: Drivers of Genome Evolution. *Science* 303, 1626-1632
- Harrison, G.P., Mayo, M.S., Hunter, E., Lever, A.M.L. 1998. Pausing of reverse transcriptase on retroviral RNA templates is influenced by secondary structures both 5' and 3' of the catalytic site. *Nucleic Acids Research* 26, 3433-3442.
- Hershhorn, A., hizi, A. 2010. Retroviral Reverse Transcriptases. *Cellular and Molecular Life Sciences* 69, 2717-2747

- Houts, G.E., Miyagi, M., Ellis, C., Beard, D., Beard, J.W. 1979. Reverse Transcriptase from Avian Myeloblastosis virus. *Journal of Virology* 29, 517-522
- Huber, H.E., Russel, M., Model, P., Richardson, C.C. 1986. Interaction of Mutant Thioredoxins of *Escherichia coli* with the Gene 5 Protein of Phage T7. *The Journal of Biological Chemistry* 261, 15006-15012.
- Ji, J., Loeb, L.A. 1992. Fidelity of HIV-1 Reverse Transcriptase Copying RNA in Vitro. *Biochemistry* 31, 954-958
- Klasens, B.I.F., Huthoff, H.T., Das, A.T., Jeeninga, R.E., Berkhout, B. 1999. The effect of template RNA structure on elongation by HIV-1 reverse transcriptase. *Biochimica et Biophysica Acta* 1444, 355-370.
- Kotewicz, M.L., Sampson, C.M., D'Alessio, J.M., Gerard, F. 1988. Isolation of cloned Moloney murine leukaemia virus reverse transcriptase lacking ribonuclease H activity. *Nucleic Acids Research* 16, 265-277
- Lien, T., Madsen, M., Rainey, F.A., Birkeland, N.K. 1998. *Petrotoga mobilis* sp. Nov., from a North sea oil-production well. *International Journal of Systematic Bacteriology* 48, 1007-1013.
- Ling, M.M., Robinson, B.H. 1997. Approaches to DNA Mutagenesis: An Overview. *Analytical Biochemistry* 254, 157-178.
- Malboeuf, C.M., Isaacs, S.J., Tran, N.H., Kim, B. 2001. Thermal Effects on Reverse Transcription: Improvement of Accuracy and Processivity in cDNA synthesis. *BioTechniques* 30,1074-1084.
- McAfee, J.G., Edmondson, S.P., Datta, P.K., Shriver, J.W., Gupta, R. 1995. Gene Cloning, Expression, and Characterisation of the Sac7d Proteins from the Hyperthermophile *Sulfolobus acidocaldarius*. *Biochemistry* 34, 10063-10077.
- Moniwa, M., Clavijo, A., Li, M., Collingnon, B., Kitching, P.R. 2007. Performance of a foot-and-mouth disease virus reverse transcription-polymerase chain reaction with amplification controls between three real-time instruments. *Journal of Veterinary Diagnostic Investigation* 19, 9-20.

- Myers, T.W., Gelfand, D.H. 1991. Reverse Transcription and DNA amplification by a *Thermus thermophilus* DNA polymerase. *Biochemistry* 30, 7661-7666
- Nakamura, T.M., Cech, T.R. 1998. Reverse Time: Origin of Telomerase. *Cell* 92, 587-590
- Nakamura, T.M., Morin, G.B., Chapman, K.B., Weinrich, S.L., Andrews, W.H., Linger, J., Harley, C.B., Cech, T.R. 1997. Telomerase Catalytic Subunit Homologs from Fission Yeast and Human. *Science* 277, 955-959
- Nazina, T.N., Tourova, T.P., Poltaraus, A.B., Novikova, E.V., Grigoryan, A.A., Ivanova, A.E., Lysenko, A.M., Petrunyaka, V.V., Osipov, G.A., Belyaev, S.S., Ivanov, M.V. 2001. Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus, Bacillus thermocatenulatus, Bacillus thermoleovorans, Bacillus kaustophilus, Bacillus thermoglucosidasius* and *Bacillus thermodenitrificans* to *Geobacillus* as the new combinations of *G. stearothermophilus, G. thermocatenulatus, G. thermoleovorans, G. kaustophilus, G. thermoglucosidasius and G. thermodenitrificans. International Journal of Systematic and Evolutionary Microbiology* 51, 433-446.
- NEB UK. DNase I. Available from http://www.neb.uk.com/productcatalogue/productinfotransfer.aspx?id=/New%20England%20Biolabs/DNA%20Modifying%20Enzymes%20and%20Cloning/Nucleases/M0303 [Accessed 3rd November 2010]
- Ng, B., Nayak, S., Gibbs, M.D., Lee, J., Bergquist, P.L. 2007. Reverse Transcriptases: Intron-encoded proteins found in thermophilic bacteria. *Gene* 393, 137-144.
- Pasloske, B.L., William, W. 1998. *Methods and Reagents for Inactivating Ribonucleases RNase A, RNaseI and RNase T1.* US patent application No. 09160284. 1998-09-24

- Pavlov, A.R., Belova, G.I., Kozyavkin, S.A., Slesarev, A.I. 2002. Helix-Hairpin-Helix motifs confer salt resistance and processivity on chimeric DNA polymerases. *PNAS*. 99, 13510-13515
- Perales, C., Cava, F., Meijer, W.J.J., Berenguer, J. 2003. Enhancement of DNA, cDNA synthesis and fidelity at high temperatures by a dimeric single-stranded DNA-binding protein. *Nucleic Acids Research* 31, 6473-6480.
- Recorbet, G., Robert, C., Givaudan, A., Kudla, B., Normand, P., and Faurie, G. 1993. Conditional Suicide System of *Escherichia coli* Released into Soil that uses the *Bacillus subtilis sacB* Gene. *Applied and Environmental Microbiology* 59, 1361-1366
- Rho, H.M., Grandgenett, D.P., Green, M. 1975. Sequence Relatedness between the Subunits of Avian Myeloblastosis Virus Reverse Transcriptase. *The Journal of Biological Chemistry* 250, 5278-5280.
- Richert, J., Kranz, E., Lörz, H., Dresselhaus, T. 1996. A reverse transcriptase-polymerase chain reaction assay for gene expression studies at the single cell level. *Plant Science* 114, 93-99.
- Rose, T.M., Schultz, E.R., Henikoff, J.G., Peitrokovski, S., McCallum, C.M., Henikoff, S. 1998. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Research* 26, 1628-1635.
- Rothschild, L.J., Mancinelli 2001. Life in Extreme Environments. *Nature* 409, 1092-1101.
- Sabot, F., Schulman, A.H. 2006. Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. *Hereditary* 97, 381-388.
- Saiki, R.K., Scharf, S., Daloona, F., Mullis, K.B., Horn, G.T., Elrich, H.A., Arnheim, N. 1985. Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anaemia. *Science* 230, 1350-1354,

- Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., Erlich, H.A. 11988. Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. *Science* 239, 487-491
- Saldanha, R., Mohr, G., Belfort, M., Lambowitz, A.M. 1993. Group I and Group II Introns. *The FASEB Journal* 7, 15-24
- Sambrook., Fritsh., Maniatis., 1989. *Molecular Cloning; A laboratory manual*. 2nd edition. USA: Cold Spring Harbour Press.
- Schwabe, W., Lee, J.E., Mathan, M., Xu, R.H., Sitaraman, K., Smith, M., Potter, R.J., Rosenthal, K., Rashtichian, A., Gerart, G.F., 1998. ThermoScript™RT, A New Avian Reverse Transcriptase For High-Temperature cDNA Synthesis To Improve RT-PCR. *FOCUS*. 20, 30-33
- Sharp, P.A. 1994. Split Genes and RNA splicing. *Cell* 77, 805-815
- Sokolova, T.G., González, J.M., Kistrikina, N.A., Chernyh, N.A., Slepova, T.V., Bonch-Osmolovskaya, E.A., Robb, F.T. 2004. *Thermosinus carboxydivorans* gen. Nov., sp. Nov., a new anaerobic, thermophilic, carbon-monoxide-oxidising, hydrogenogenic bacterium from a hot pool of Yellowstone National Park. *International Journal of Systematic and Evolutionary Microbiology* 54, 2353-2359.
- Tabor, S., Huber, H.E., Richardson, C.C. 1987. *Escherichia coli* Thioredoxin Confers Processivity on the DNA Polymerase Activity of the Gene 5 Protein of Bacteriophage T7. *The Journal of Biological Chemistry* 22, 16212-16223.
- Takano, T., Miyauchi, A., Matsuzuka, F., Yoshida, H., Kuma, K., Amino, N. 2000. Diagnosis of thyroid malignant lymphoma by reverse transcription-polymerase chain reaction detecting the monoclonality of immunoglobulin heavy chain messenger ribonucleic acid. *The Journal of Clinical Endocrinology and Metabolism* 85, 671-675
- Takami, H., Nishi, S., Lu, J., Shimamura, S., Takaki, Y. 2004. Genomic characterisation of thermophilic *Geobacillus* species isolate from the deepest sea mud of the Mariana Trench. *Extremophiles* 8, 351-356

- Temin, H.M. 1970. RNA-Dependent DNA Polymerase in Virions of Rous Sarcoma Virus. *Nature* 226, 1211-1213
- Toro, N. 2003. Bacteria and Achaea Group II Introns: additional mobile genetic elements in the environment. *Environmental Microbiology* 5, 143-151
- Toro, N., Jiménez-Zurdo, J.I., García-Rodríguez. 2007. Bacterial Group II Introns: not just splicing. *FEMS Microbiology Reviews*. 31, pp. 342-358.
- Towner, J.S., Rollin, P.E., Bausch, D.G., Sanchez, A., Crary, S.M., Vincent, M., Lee, W.F., Spiropoulou, C.F., Ksiazek, T.G., Lukwiya, M., Kaducu, F., Downing, R., Nichol, S.T. 2004. Rapid Diagnosis of Ebola Hemorrhagic Fever by Reverse Transcription-PCR in an Outbreak Setting and Assessment of Patient Viral Load as a Predictor of Outcome. *Journal of Virology* 78, 4330-4341.
- Travisano, M., Inouye, M. 1995. Retrons: retroelements of no known function. *Trends in Microbiology* 3, 209-211.
- Triglia, T., Peterson, M.G., Kemp, D.J. 1988. A procedure for *in vitro* amplification of DNA segments that lie outside the boundaries of known sequences. *Nucleic Acids Research* 16, 8186.
- Vellore, J., Moretz, S.E., Lampson, B.C., 2004. A Group II Intron-Type Open Reading Frame from the Thermophile *Bacillus* (*Geobacillus*) *stearothermophilus* Encodes a Heat-stable Reverse Transcriptase. *Applied and Environmental Microbiology* 70 , 7140-7147.
- Verma, I.M. 1975. Studies on Reverse Transcriptase of RNA Tumour Viruses III. Properties of Purified Moloney Murine Leukaemia Virus DNA polymerase and Associated RNase H. *Journal of Virology* 15, 843-854
- Wang, Y., Prosen, D.E., Mei, L., Sullivan, J.C., Finney, M., Vander Horn, P.B. 2004. A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance *In vitro*. *Nucleic Acids Research* 32, 1197-1207.

- Wank, H., SanDilippo, J., Singh, E.N., Matsuura, M., Lambowitz, A.M. 1999. A Reverse Transcriptase/Maturase Promotes Splicing by Binding at Its Own Coding Segment in a Group II Intron. *Molecular Cell.* 4, 239-250
- Wilhelm, M., and Wilhelm, F.-X. 2001. Reverse transcription of retroviruses and LTR retrotransposons. *Cellular and Molecular Life Sciences* 58, 1246-1262.
- Xiong, Y and Eickbush, T.H. 1990. Origin and Evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO Journal*. 9, 3353-3362
- Zale, S.E., Klibanov, M. 1986. Why Does Ribonuclease Irreversibly Inactivate at High Temperatures. *Biochemistry* 25, 5432-5444.
- Zimmerly, S., Moran, J.V., Perlman, P.S., Lambowitz, A.M. 1999. Group II Intron Reverse Transcriptase in Yeast Mitochondria. Stabilisation and Regulation of Reverse Transcriptase Activity by the Intron RNA. *Journal of Molecular Biology* 289, 473-490.

APPENDIX

I – Primers

16s Primers

27F: 5′ aga gtt tga tcm tgg ctc ag 3′

1429R: 5′ tac ggy tac ctt gtt acg act t 3′

CODEHOP primers

RTF1: 5′ cac aaa act gag gaa ggn can ccn car gg 3′

RTF2: 5′ cac aaa act gag gaa ggn gtn ccn car gg 3′

RTR1: 5′ gac ggt gat tac gaa rtc rtc cgc rat 3′

Screening primers

pCR®2.1 and pUC screening primers

M13 forward: 5′ AGC GGA TAA CAA TTT CAC ACA GGA 3′

M13 Reverse: 5′ GCG GTC CCA AAA GGG TCA GTG CTG 3′

pET vector screening primers

T7 Promoter: 5′ AAA TTA ATA CGA CTC ACT ATA GGG 3′

T7 Terminator: 5′ GCT AGT TAT TGC TCA GCG G 3′

pUC19 screening primers

pUC19_F_Screen: 5′ atg ttg aat act cat act ctt cct 3′

pUC19_R_Screen: 5′ tta cag aca agc tgt gac cg 3′

Gene-Walking primers – *B. caldovelox*

B.cv_GW1_NterR_nest: 5′ gtg gac tga gcg gcc c 3′

B.cv_GW1_NterR: 5′ cag gag aat gtt gga cag g 3′

B.cv_GW1_CterF: 5′ gag ctg gac aaa gaa ttg g 3′

B.cv_GW1_CterF_Nest: 5′ aac gag ggc aca agt ttg 3′

Gene-Walking Primers – *B. caldotenax*

B.ctnx_GW1_NterR_Nest: 5′ cag cag cgg tga gag tat G 3′

B.ctnx_GW1_NterR: 5′ acc gtc aga gcg atg ttg 3′
B.ctnx_GW1_CterF: 5′ ccg tgt gaa atg gga tg 3′
B.ctnx_GW1_CterF_nest: 5′ atc gga cat ctc agc gg 3′

Gene-Walking Primers – New Zealand *B. Stearothermophilus* Strain 11

NZ_GW1_NterR_Nest: 5′ agg ggg ctg agg gga ccg 3′
NZ_GW1_CterF_Nest: 5′ cgt gga ttg aaa ttc tgc cgc 3′
NZ_Gw1_CterF: 5′ tta gac aag gag ttg gag aag 3′
NZ_GW1_NterR: 5′ cga gaa gga tgt tcg cc 3′

NZ_GW2_CterF_nest: 5′ ctg aaa cag cgg att cga c 3′
NZ_GW2_CterR_nest: 5′ ttt cac ata gat gtt gca 3′
NZ_GW2_CterF: 5′ gga acg aaa agc gcg aat c 3′
NZ_GW2_CterR: 5′ ctt tgt ttc acc cgt tgc c 3′

7.6_GW3_Cterf_Nest: 5′ ata ccc gaa aag gag ctt ggc 3′
7.6_GW3_CterF: 5′ gga gac agc ggt gat gga gat c 3′
7.6_GW3_CterR: 5′ tca cga att ctg gtt ctg acc c 3′
7.6_GW3_CterR_Nest: 5′ ttt cca ttg aag cca ttg aca gag 3′

Cloning primers – *B. caldovelox*

Bcv_IEP_F1_Start: 5′ gct tag cat atg tgg atg aaa cag gta atg tca cga 3′
B.cv_IEP_R1_EcoRI: 5′ ttg tct gga att cgt ttc gct aat atc c3′
B.cv_IEP_F2_EcoRI: 5′ gcg aaa cga att cca gac aaa atc 3′
B.cv_IEP_R2_Stop_SalI: 5′ aag ctt gga tcc gtc gac tta agt ctg acg cag aga ttc ata tcg 3′

Cloning Primers – *Thermosinus carboxydivorans*

T.c_IEP_F_NdeI: 5′ cac gaa tta cat atg atg gca tcc cga ccg acc 3.
T.c_IEP_R_XhoI: 5′ gaa tcc gtc gac ctc gag tta acc acc gga cgt gcg 3′
T.c_IEP_R_XhoI_His: 5′ gaa tcc gtc gac ctc gag cac cac cgg acg tgc g 3′

Cloning Primers – *Petratoga mobilis*

P.m_IEP1/2_F_NdeI: 5′ gga ggt ttc gac ata tga aag atg cga agg 3′

P.m_IEP3_F_NdeI: 5′ gga ggt ttc gac ata tga aag atg cga aga 3′

P.m_IEP1_R_XhoI: 5′ ccg ttc ggt ata cct cga gtt aac aga att cat gc 3′

P.m_IEP2_R_XhoI: 5′ ccg ttc gga ata cct cga gtt aac aga att gat gt 3′

P.m_IEP3_R_XhoI: 5′ ccg ttc ggt ata cct cga gtt aac aga att gat gt 3′

New Zealand strain 11 *B. stearothermophilus*

NZ11_F_NdeI: 5′ GAG GAA GAG CAT ATG GCT TTG TTG GAA CGC ATC 3′

NZ11_R_XhoI: 5′ GAG GCA TTC CTC GAG ATT AAC CTT GAC GGA GTT CGA AAT AT 3′

Site Directed Mutagenesis Primers

Bcv_NdeI_SDM_F: 5′ gag tgc cca agc aga aag c**c**t atg aat ggg gaa aca c 3′

Bcv_NdeI_SDM_R: 5′ gtg ttt ccc cat tca tag gct ttc tgc ttg ggc act c 3′

Overlap Extension Primers

*P1 - Sac7d _Nter_F_*NdeI: 5′ ccg tcc ccg cat atg gtg aaa gtg aaa ttt aaa tac aaa ggt g 3′

P2 - Sac7d _Nter_R: 5′ ctc gtg aca tta cct gtt tca tcc aca taa cgg tac cac ctt ttt tct cgc gtt ccg 3′

P3 - B.cv_Cter_F: 5′ cgt gcg gaa cgc gag aaa aaa ggt ggt acc gtt atg tgg atg aaa cag gta atg tca cga g

*P4 - Bcv_IEP_Cter_R_*XhoI: 5′ ccg ccg ctc gag att aag tct gac gca gag att cat atc g 3′

*P5 - Bcv_IEP_Nter_F*_NdeI: 5′ ccg tcc ccg cat atg tgg atg aaa cag gta atg tca c 3′

P6: B.cv_Nter_R: 5′ acc ttt gta ttt aaa ttt cac ttt cac cat aac ggt acc agt ctg acg ca gaga ttc ata tcg tt 3′

P7: Sac7d _Cter_F: 5′ tat caa cga tat gaa tct ctg cgt cag act ggt acc gtt atg gtg aaa gtg aaa ttt aaa tac aaa ggt 3′

*P8: Sac7d _Cter_R*_XhoI: 5′ ccg ccg ctc gag att aac ctt ttt tct cgc gtt ccg 3′

Overlap Extension Primers - To remove GTV linker

Sac7_Nter_L_No linker: 5′ ctc gtg aca tta cct gtt tca tcc aca tac ctt ttt tct cgc gtt ccg cac g 3′

Sac7_Nter_U_No linker: 5′ cgt gcg gaa cgc gag aaa aaa ggt atg tgg atg aaa cag gta atg tca cga g 3′

Reverse Transcription primers

MS2:3014_F: 5′ agg tat gtc aga tcc acg cct cta t 3′

MS2:2475_F: 5′ tcg atc aac cag cgt ctg gct cag c 3′

MS2:3231_F: 5′ ttc agc gaa aag cac gac agt ggt c 3′

MS2:13_F: 5′ ttt cgg ggt cct gct caa ctt cct g 3′

MS2:1868_F: 5′ tcc gca cag tga cga ctt tac agc a 3′

MS2:3529_R: 5′ aat ccc ggg tcc tct ctt tag ggg g 3′

MS2:3193_R: 5′ cga gcg ata cga gca aga cgg aaa c 3′

MS2:3127_R: 5′ ttg gtg tat acc gag act gcc gta g 3′

MS2:3395_R: 5′ tta cgg ggg tcc ctc ggt cag cta c 3′

RT-qPCR primers

MS2:3231_F: 5′ ttc agc gaa aag cac gac agt ggt c 3′

MS2:3395_R: 5′ tta cgg ggg tcc ctc ggt cag cta c 3′

28S_F: 5′ aac gag att ccc act gtc cct a 3′

28S_R: 5′ ggt ctt ctt tcc ccg ctg att 3′

Mitochondrial_ATP_Synthase_F: 5′ cag tga tta tag gct ttc gct cta a 3′

Mitochondrial_ATP_Synthase_R: 5′ cag ggc tat tgg ttc aat gag ta 3′

GAPDH_F: 5′ gac agt cag ccg cat ctt ctt 3′

GAPDH_R: 5′ tcc gtt gac tcc gac ctt ca 3′

β-2-microglobulin_F: 5′ aag gac tgg tct ttc tat ctc ttg ta 3′

β-2-microglobulin_R: 5′ act taa cta tct tgg gct gtc aca 3′

Fidelity Assay

LacI^{q}_F_SacII: 5′ gaa ttc cat atc ccg cgg ata atg agg ttc tgt acc cga cac 3′

LacI^{q}_R_SacII: 5′ gaa ttc cat atg ccg cgg ctg acg tga ggt tct tgc 3′

LacI^{q}_U_seq: 5′ cgt ggt ggt gtc gat ggt ag 3′

LacI^{q}_F_seq: 5′ gtt gca gca agc ggt cca cg 3′

SacB_F_AatII: 5′ tta cca gac gtc cat atg cct gcc gtt cac t 3′

SacB_R_AatII: 5′ gaa ttc gac gtc gga tcc cgg cat ttt ctt3′

DNA-Dependent DNA Polymerase Activity

-40M13LFF: 5′ Fam-gtt ttc cca gtc acg acg ttg taa aac gac ggc c 3′

RNA-Dependent DNA Polymerase Activity

FAM_MS2:3193R: 5′ Fam-cga gcg ata cga gca aga cgg aaa c3′

II – Genotype of *E. coli* Strains

TOP10F'
F´{*lac*Iq, Tn*10*(TetR)} *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara leu*) 7697 *gal*U *gal*K *rps*L (StrR) *end*A1 *nup*G

TOP10
F– *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara leu*) 7697 *gal*U *gal*K *rps*L (StrR) *end*A1 *nup*G

KRX
[F2, *tra*D36, ⊗*omp*P, *pro*A+B+, *lacI*q, ⊗(*lac*Z)M15] ⊗*omp*T, *end*A1, *rec*A1, *gyr*A96 (Nalr), *thi*-1, *hsd*R17 (rk–, mk+), e14– (McrA–), *rel*A1, *sup*E44, ⊗(*lac-pro*AB), ⊗(*rha*BAD)::T7 RNA polymerase

ArcticExpress™
E. coli B F– ompT hsdS(rB– mB–) dcm+ Tetr gal λ(DE3) endA Hte [cpn10 cpn60 Gentr] [argU ileY leuW Strr]

III – DNA and Protein Sequences

Geobacillus kaustophilus recombinase of Bh.Int-like element DNAsequence

```
   1 atgtggatga aacaggtaat gtcacgagag aatctcctgc gagcactcaa
  51 acaagtggaa aagaataaag ggtcccatgg aaccgatgga atgtccgtca
 101 aagacctgcg aagacacctc gtggaacatt gggacgcgat acggcacgct
 151 ttagaagaag ggacctacga accttgcccg gtccgacggg tcgaaatccc
 201 gaaaccgaac ggaggagtca ggttactagg aatcccgacc gtgacagacc
 251 ggttcatcca acaggccatc gcccaagtgc tcacgccgat ctttgaccca
 301 tccttttcgg aacacagcta cgggtttcgt cccggtcgaa gaggacacga
 351 cgcggtgaaa aaggcgaagc agtatattca ggaaggatat acatgggtgg
 401 tagatatcga cttggaaaag ttctttgatc gagtcaacca tgacaaactg
 451 atggggatat tagcgaaacg aattccagac aaaatcctcc taaagttgat
 501 acggaagtat ctacaggcag gggtcatgat caacggggtg gtcatggaaa
 551 cacaagaggg gactccacaa ggagggccgc tcagtccact cctgtccaac
 601 attctcctgg atgagctgga caaagaattg gaaaaacgag ggcacaagtt
 651 tgtacggtat gcggatgact gcaatatcta cgtaaggacg aagaaggccg
 701 gggaacgggt gatgaaatcg atcacggcat tcatcgaaaa gaaactccgg
 751 ctgaaagtca acgaaaccaa atcggcagtg gatcgtccgt ggaggagaaa
 801 attcctcggt tttagcttca ccccaagtaa ggagccaaaa atccgaattg
 851 caaaggaaag cattcggcgc atgaagcaaa ggatacgcac catgacgagc
 901 cgatcgaaac cgattcccat gcccgaacga atcgaacagc tcaatcaata
 951 cattctggga tggtgtggat acttttcgct agcagagacc ccaagtgtgt
1001 tcaaagaact agatggatgg attcgacgaa ggctgcgcat gtgccaatgg
```

```
1051 aaagagtgga aacttccgag aaccagagtc cgaaaactgc aaagtttagg
1101 agtgcccaag cagaaagcat atgaatgggg aaacactcgg aagaaatatt
1151 ggagagtggc cgccagtccc atcctgcata aagcccttgg caactcctat
1201 tgggagagcc aagggctgaa gagtctttat caacgatatg aatctctgcg
1251 tcagacttaa
```

The highlighted region represents the sequence that was identified during gene walking of *B. caldovelox.*

Geobacillus kaustophilus recombinase of Bh.Int-like element protein sequence

```
  1 MWMKQVMSRE NLLRALKQVE KNKGSHGTDG MSVKDLRRHL VEHWDAIRHA
 51 LEEGTYEPCP VRRVEIPKPN GGVRLLGIPT VTDRFIQQAI AQVLTPIFDP
101 SFSEHSYGFR PGRRGHDAVK KAKQYIQEGY TWVVDIDLEK FFDRVNHDKL
151 MGILAKRIPD KILLKLIRKY LQAGVMINGV VMETQEGTPQ GGPLSPLLSN
201 ILLDELDKEL EKRGHKFVRY ADDCNIYVRT KKAGERVMKS ITAFIEKKLR
251 LKVNETKSAV DRPWRRKFLG FSFTPSKEPK IRIAKESIRR MKQRIRTMTS
301 RSKPIPMPER IEQLNQYILG WCGYFSLAET PSVFKELDGW IRRRLRMCQW
351 KEWKLPRTRV RKLQSLGVPK QKAYEWGNTR KKYWRVAASP ILHKALGNSY
401 WESQGLKSLY QRYESLRQT
```

Length in amino acids: 419
Molecular weight in kD: 49.09
Iso-electric point (IP): 10.14

Sac7d *E. coli* codon optimized DNA sequence

```
  1 atggtgaaag tgaaatttaa atacaaaggt gaagagaaag aagttgacac
 51 ttctaaaatt aaaaaagtgt ggcgtgtggg taaaatggtt tctttcacct
101 atgatgataa cggcaaaacc ggtcgcggtg cggtgtctga gaaagacgcg
151 ccgaaagaac tgctggatat gctggcgcgt gcggaacgcg agaaaaaagg
201 ttaa
```

Sac7d *E. coli* codon optimized protein sequence

```
  1 MVKVKFKYKG EEKEVDTSKI KKVWRVGKMV SFTYDDNGKT GRGAVSEKDA
 51 PKELLDMLAR AEREKKG
```

Sac7d-*B.cv* IEP fusion gene sequence

```
  1 atggtgaaag tgaaatttaa atacaaaggt gaagagaaag aagttgacac
 51 ttctaaaatt aaaaaagtgt ggcgtgtggg taaaatggtt tctttcacct
101 atgatgataa cggcaaaacc ggtcgcggtg cggtgtctga gaaagacgcg
```

```
 151 ccgaaagaac tgctggatat gctggcgcgt gcggaacgcg agaaaaaagg
 201 tggtaccgtt atgtggatga aacaggtaat gtcacgagag aatctcctgc
 251 gagcactcaa acaagtggaa aagaataaag ggtcccatgg aaccgatgga
 301 atgtccgtca aagacctgcg aagacacctc gtggaacatt gggacgcgat
 351 acggcacgct ttagaagaag ggacctacga accttgcccg gtccgacggg
 401 tcgaaatccc gaaaccgaac ggaggagtca ggttactagg aatcccgacc
 451 gtgacagacc ggttcatcca acaggccatc gcccaagtgc tcacgccgat
 501 ctttgaccca tccttttcgg aacacagcta cgggtttcgt cccggtcgaa
 551 gaggacacga cgcggtgaaa aaggcgaagc agtatattca ggaaggatat
 601 acatgggtgg tagatatcga cttggaaaag ttctttgatc gagtcaacca
 651 tgacaaactg atggggatat tagcgaaacg aattccagac aaaatcctcc
 701 taaagttgat acggaagtat ctacaggcag gggtcatgat caacggggtg
 751 gtcatggaaa cacaagaggg gactccacaa ggagggccgc tcagtccact
 801 cctgtccaac attctcctgg atgagctgga caaagaattg gaaaaacgag
 851 ggcacaagtt tgtacggtat gcggatgact gcaatatcta cgtaaggacg
 901 aagaaggccg gggaacgggt gatgaaatcg atcacggcat tcatcgaaaa
 951 gaaactccgg ctgaaagtca acgaaaccaa atcggcagtg gatcgtccgt
1001 ggaggagaaa attcctcggt tttagcttca ccccaagtaa ggagccaaaa
1051 atccgaattg caaaggaaag cattcggcgc atgaagcaaa ggatacgcac
1101 catgacgagc cgatcgaaac cgattcccat gcccgaacga atcgaacagc
1151 tcaatcaata cattctggga tggtgtggat acttttcgct agcagagacc
1201 ccaagtgtgt tcaaagaact agatggatgg attcgacgaa ggctgcgcat
1251 gtgccaatgg aaagagtgga aacttccgag aaccagagtc cgaaaactgc
1301 aaagtttagg agtgcccaag cagaaagcct atgaatgggg aaacactcgg
1351 aagaaatatt ggagagtggc cgccagtccc atcctgcata aagcccttgg
1401 caactcctat tgggagagcc aagggctgaa gagtctttat caacgatatg
1451 aatctctgcg tcagacttaa
```

Sac7d-*B.cv* IEP fusion protein sequence

```
   1 MVKVKFKYKG EEKEVDTSKI KKVWRVGKMV SFTYDDNGKT GRGAVSEKDA
  51 PKELLDMLAR AEREKKGGTV MWMKQVMSRE NLLRALKQVE KNKGSHGTDG
 101 MSVKDLRRHL VEHWDAIRHA LEEGTYEPCP VRRVEIPKPN GGVRLLGIPT
 151 VTDRFIQQAI AQVLTPIFDP SFSEHSYGFR PGRRGHDAVK KAKQYIQEGY
 201 TWVVDIDLEK FFDRVNHDKL MGILAKRIPD KILLKLIRKY LQAGVMINGV
 251 VMETQEGTPQ GGPLSPLLSN ILLDELDKEL EKRGHKFVRY ADDCNIYVRT
 301 KKAGERVMKS ITAFIEKKLR LKVNETKSAV DRPWRRKFLG FSFTPSKEPK
 351 IRIAKESIRR MKQRIRTMTS RSKPIPMPER IEQLNQYILG WCGYFSLAET
 401 PSVFKELDGW IRRRLRMCQW KEWKLPRTRV RKLQSLGVPK QKAYEWGNTR
 451 KKYWRVAASP ILHKALGNSY WESQGLKSLY QRYESLRQT
```

Length in amino acids: 489
Molecular weight in kD: 57.00
Iso-electric point (IP): 10.06

B.cv IEP-Sac7d fusion gene sequence

```
   1 atgtggatga aacaggtaat gtcacgagag aatctcctgc gagcactcaa
  51 acaagtggaa aagaataaag ggtcccatgg aaccgatgga atgtccgtca
```

```
 101 aagacctgcg aagacacctc gtggaacatt gggacgcgat acggcacgct
 151 ttagaagaag ggacctacga accttgcccg gtccgacggg tcgaaatccc
 201 gaaaccgaac ggaggagtca ggttactagg aatcccgacc gtgacagacc
 251 ggttcatcca acaggccatc gcccaagtgc tcacgccgat ctttgaccca
 301 tccttttcgg aacacagcta cgggtttcgt cccggtcgaa gaggacatga
 351 cgcggtgaaa aaggcgaagc agtatattca ggaaggatat acatgggtgg
 401 tagatatcga cttggaaaag ttctttgatc gagtcaacca tgacaaactg
 451 atggggatat tagcgaaacg aattccagac aaaatcctcc taaagttgat
 501 acggaagtat ctacaggcag gggtcatgat caacggggtg gtcatggaaa
 551 cacaagaggg gactccacaa ggagggccgc tcagtccact cctgtccaac
 601 attctcctgg atgagctgga caaagaattg gaaaaacgag ggcacaagtt
 651 tgtacggtat gcggatgact gcaatatcta cgtaaggacg aagaaggccg
 701 gggaacgggt gatgaaatcg atcacggcat tcatcgaaaa gaaactccgg
 751 ctgaaagtca acgaaaccaa atcggcagtg gatcgtccgt ggaggagaaa
 801 attcctcggt tttagcttca ccccaagtaa ggagccaaaa atccgaattg
 851 caaaggaaag cattcggcgc atgaagcaaa ggatacgcac catgacgagc
 901 cgatcgaaac cgattcccat gcccgaacga atcgaacagc tcaatcaata
 951 cattctggga tggtgtggat acttttcgct agcagagacc ccaagtgtgt
1001 tcaaagaact agatggatgg attcgacgaa ggctgcgcat gtgccaatgg
1051 aaagagtgga aacttccgag aaccagagtc cgaaaactgc aaagtttagg
1101 agtgcccaag cagaaagcct atgaatgggg aaacactcgg aagaaatatt
1151 ggagagtggc cgccagtccc atcctgcata aagcccttgg caactcctat
1201 tgggagagcc aagggctgaa gagtctttat caacgatatg aatctctgcg
1251 tcagactggt accgttatgg tgaaagtgaa atttaaatac aaaggtgaag
1301 agaaagaagt tgacacttct aaaattaaaa aagtgtggcg tgtgggtaaa
1351 atggtttctt tcacctatga tgataacggc aaaaccggtc gcggtgcggt
1401 gtctgagaaa gacgcgccga aagaactgct ggatatgctg gcgcgtgcgg
1451 aacgcgagaa aaaaggttaa
```

B.cv IEP-Sac7d fusion protein sequence

```
   1 MWMKQVMSRE NLLRALKQVE KNKGSHGTDG MSVKDLRRHL VEHWDAIRHA
  51 LEEGTYEPCP VRRVEIPKPN GGVRLLGIPT VTDRFIQQAI AQVLTPIFDP
 101 SFSEHSYGFR PGRRGHDAVK KAKQYIQEGY TWVVDIDLEK FFDRVNHDKL
 151 MGILAKRIPD KILLKLIRKY LQAGVMINGV VMETQEGTPQ GGPLSPLLSN
 201 ILLDELDKEL EKRGHKFVRY ADDCNIYVRT KKAGERVMKS ITAFIEKKLR
 251 LKVNETKSAV DRPWRRKFLG FSFTPSKEPK IRIAKESIRR MKQRIRTMTS
 301 RSKPIPMPER IEQLNQYILG WCGYFSLAET PSVFKELDGW IRRRLRMCQW
 351 KEWKLPRTRV RKLQSLGVPK QKAYEWGNTR KKYWRVAASP ILHKALGNSY
 401 WESQGLKSLY QRYESLRQTG TVMVKVKFKY KGEEKEVDTS KIKKVWRVGK
 451 MVSFTYDDNG KTGRGAVSEK DAPKELLDML ARAEREKKG
```

Length in amino acids: 489
Molecular weight in kD: 57.00
Iso-electric point (IP): 10.06

Thermosinus carboxydivorans IEP gene sequence

```
   1 ttgatggcat cccgaccgac cggcttcggg accaaatacg cgccgaatgg
  51 ccgcgcatcc gggaagaact gctcgcgggt acctacaaac cgatgcccgt
 101 gcgccgggtc gaaatcccga aagtgcggag gaggtacacg gatgctgggg
 151 atacccaccg tgatggaccg cctgatccag caggcccttc tgcaggtact
 201 gacgcatatc tttgacccgc acttttccga agccagctac gggttccgtc
 251 ctggcaagaa agcacatgat gcggtaagga aggcgcgcca atacgtggaa
 301 gaaggctatg aatgggctgt ggacatggac cttgagaaat tcttcgacag
 351 ggtaaaccat gatatactca tggcccgagt ggcccgaaaa gtaacagaca
 401 aaagagtgtt aaaactcatc cgccgctatc tccaggcagg catcatggtc
 451 aatggtgtgg tcatggacag ggaagaagga acaccgcaaa gcggaccatt
 501 aagcccactc ctggccaaca tcctactgga tgacctggat aaggaactgg
 551 agaaacgtgg ccacaagttc atgcgttatg ccgatgactg caatatctac
 601 gtaaagacta ggcgcgcggg tgaaagaata ttgactagtg tccgcaacta
 651 cttgcaggag cggttaaaac tcaaactgaa cgaagagaaa agcatggtag
 701 accgaccgtg gaaactgaaa tttctgggct ttagcatgta taaagccaaa
 751 ggtgggaaaa tcctcatccg cctggcgtca caaacaatcg accgggtgaa
 801 acagaaaatc cgggaagcga ctgctcgtag tgctccacaa tcaatggcgg
 851 agcggataga acgcctgaac acctatttgg gaggatggat aggatacttc
 901 gccttggcgg atactcccag cgtctttaag aacatagacg gctggatacg
 951 gagaagatta cgcatgtgcc tgtggaagca gtggaagcga gtgaggacca
1001 gatacagtga gttaagggct ttgggactac cggaatgggt agtgcatgaa
1051 ttcgccaata cccgcaaagg accatggcga atggcccacg ggccaatgaa
1101 tagagccctg ggcaatgcct actggcgagc ccagggcctg atgagtttaa
1151 ccgaacgtta ccaaaagctt cgccaagctt ggcgaaccgc cggatgcgga
1201 cccgcacgtc cggtggtgtg a
```

Thermosinus carboxydivorans IEP sequence

```
   1 LMASRPTGFG TKYAPNGRAS GKNCSRVPTN RCPCAGSKSR KCGGGTRMLG
  51 IPTVMDRLIQ QALLQVLTHI FDPHFSEASY GFRPGKKAHD AVRKARQYVE
 101 EGYEWAVDMD LEKFFDRVNH DILMARVARK VTDKRVLKLI RRYLQAGIMV
 151 NGVVMDREEG TPQSGPLSPL LANILLDDLD KELEKRGHKF MRYADDCNIY
 201 VKTRRAGERI LTSVRNYLQE RLKLKLNEEK SMVDRPWKLK FLGFSMYKAK
 251 GGKILIRLAS QTIDRVKQKI REATARSAPQ SMAERIERLN TYLGGWIGYF
 301 ALADTPSVFK NIDGWIRRRL RMCLWKQWKR VRTRYSELRA LGLPEWVVHE
 351 FANTRKGPWR MAHGPMNRAL GNAYWRAQGL MSLTERYQKL RQAWRTAGCG
 401 PARPVV
```

Length in amino acids: 406
Molecular weight in kD: 46.74
Iso-electric point (IP): 10.35

Petrotoga mobilis SJ95 IEP1 Gene Sequence

```
   1 ttgaaagatg cgaagggaca cggaaaaagc atacaacttc gattagaagg
  51 tttcctacat gaagataagg gagagcctga aaataatgta gaagcgccta
 101 gtataacttc tacgtctgaa agaggaagaa acgatgataa gggatgcagt
 151 gaagggatgc ttgaaaagat attatccaag gataatatga ataaagcata
 201 caagaaggta aaggccaaca aaggagcccc tgggatagac gggatgaaag
 251 tagaagaact ctttggatat ctcagacaac acggagaaga attaaggcaa
 301 gagctccttg aaggaaggta caccccgaag tcggtaagga ggaaagaaat
 351 accgaaaccg gatggaggga aaagactact aggcatacca acatcaattg
 401 acagagtaat ccagcaatcc atagcccagg tgttgacacc tatttacgaa
 451 aagaaatttg tagataacag ttatggattc aggccattac gggatgcaaa
 501 acaagccata cgaaaaagca aagaatacct aaacaaaggg catacatggg
 551 tagtggacat agacttagaa cgctactttg acacagtcaa ccatgacaaa
 601 ctgatgagga taatatcaaa agacgtaaaa gatggcagag tgatatccct
 651 gataaggaaa tacctcaaga gtggggtaat ggtaaatggg gtagtaatag
 701 aaactgaaga agggactcca caaggggac cgctatcccc attactaagc
 751 aacataatgc tccacgaact tgacgtagaa ctaacgaaaa ggggacacaa
 801 gttttgcagg tatgcagacg attgcaacat atacgtgaaa agcgagaaat
 851 ccgcatatag ggtgatggaa agtataacaa aatacataga gaagaaatta
 901 aagctaaaag tgaatagcaa gaaaagcaaa gtagtcagac cttgggacct
 951 caaatactta gggttctcct tctacgtgaa agaagagaag tacgaaatta
1001 gagtccatgg aaaatccatt aaagaattca aaaagaagtt aaaaggagaa
1051 acgaagagaa gcagtggaag aagcatggca tacaggctgt caaggataaa
1101 acaaataata acaggatgga caaactacta tggaatagca aacatgaaat
1151 cgatagcggg aagtcttgat ggatggctta gaaggagaat taggatgtgt
1201 atctggaagg aatggaagaa aatcaaaaca aggtataaaa acttagtcaa
1251 attagggtta aacatctaca aagcatggga atatgcaaat acaaggaaag
1301 gctattggag aatctccaac agccccatac tttcaagaac actaactaat
1351 aaacacctta agaaaatggg gttaacttcg atcctggaaa cgtataacct
1401 aaagcatgaa ttctgttga
```

Petrotoga mobilis SJ95 IEP1 Sequence

```
   1 lkdakghgks iqlrlegflh edkgepennv eapsitstse rgrnddkgcs
  51 egmlekilsk dnmnkaykkv kankgapgid gmkveelfgy lrqhgeelrq
 101 ellegrytpk svrrkeipkp dggkrllgip tsidrviqqs iaqvltpiye
 151 kkfvdnsygf rplrdakqai rkskeylnkg htwvvdidle ryfdtvnhdk
 201 lmriiskdvk dgrvislirk ylksgvmvng vvieteegtp qggplsplls
 251 nimlheldve ltkrghkfcr yaddcniyvk seksayrvme sitkyiekkl
 301 klkvnskksk vvrpwdlkyl gfsfyvkeek yeirvhgksi kefkkklkge
 351 tkrssgrsma yrlsrikqii tgwtnyygia nmksiagsld gwlrrrirmc
 401 iwkewkkikt ryknlvklgl niykaweyan trkgywrisn spilsrtltn
 451 khlkkmglts iletynlkhe fc
```

Length in amino acids: 472
Molecular weight in kD: 54.74
Iso-electric point (IP): 9.86

Petrotoga mobilis SJ95 IEP2 Gene Sequence

```
   1 ttgaaagatg cgaagggaca cggaaaaagc atacaacttc gattagaagg
  51 tttcctacat gaagataagg gagagcctga aaataatgta gaagcgccta
 101 gtataacttc tacgtctgaa agaggaagaa acgatgataa gggatgcagt
 151 gaagggatgc ttgaaaagat attatccaag gataatatga ataaagcata
 201 caagaaggta aaggccaaca aaggagcccc tgggatagac gggatgaaag
 251 tagaagaact ctttggatat ctcagacaac acggagaaga attaaggcaa
 301 gagctccttg aaggaaggta caccccgaag tcggtaagga ggaaagaaat
 351 accgaaaccg gatggaggga aaagactact aggcatacca acatcaattg
 401 acagagtaat ccagcaatcc atagcccagg tgttgacacc tatttacgaa
 451 aagaaatttg tagataacag ttatggattc aggccattac gggatgcaaa
 501 acaagccata cgaaaaagca aagaatacct aaacaaaggg catacatggg
 551 tagtggacat agacttagaa cgctactttg acacagtcaa ccatgacaaa
 601 ctgatgagga taatatcaaa agacgtaaaa gatggcagag tgatatccct
 651 gataaggaaa tacctcaaga gtggggtaat ggtaaatggg gtagtaatag
 701 aaactgaaga agggactcca caagggggac cgctatcccc attactaagc
 751 aacataatgc tccacgaact tgacgtagaa ctaacgaaaa ggggacacaa
 801 gttttgcagg tatgcagacg attgcaacat atacgtgaaa agcgagaaat
 851 ccgcatatag ggtgatggaa agtataacaa aatacataga gaagaaatta
 901 aagctaaaag tgaatagcaa gaaaagcaaa gtagtcagac cttgggacct
 951 caaatactta gggttctcct tctacgtgaa agaagagaag tacgaaatta
1001 gagtccatgg aaaatccatt aaagaattca aaaagaagtt aaaaggagaa
1051 acgaagagaa gcagtggaag aagcatggca tacaggctgt caaggataaa
1101 acaaataata acaggatgga caaactacta tggaatagca aacatgaaat
1151 cggtagcggg aagtcttgat ggatggctca gaaggagaat taggatgtgt
1201 atctggaagg aatggaagaa aatcaaaaca aaacatgaaa acttagtcaa
1251 attaggacta aacacctaca aagcatggga atatgcaaac acaagaaaag
1301 gctattggag gatctccaac agcccgatac tttcaatgac actaactaat
1351 aaacgcctta aggaaatggg gttaacttcg atcctggaaa cgtataacct
1401 aaaacatcaa ttctgttga
```

Petrotoga mobilis SJ95 IEP2 Sequence

```
   1 lkdakghgks iqlrlegflh edkgepennv eapsitstse rgrnddkgcs
  51 egmlekilsk dnmnkaykkv kankgapgid gmkveelfgy lrqhgeelrq
 101 ellegrytpk svrrkeipkp dggkrllgip tsidrviqqs iaqvltpiye
 151 kkfvdnsygf rplrdakqai rkskeylnkg htwvvdidle ryfdtvnhdk
 201 lmriiskdvk dgrvislirk ylksgvmvng vvieteegtp qggplsplls
 251 nimlheldve ltkrghkfcr yaddcniyvk seksayrvme sitkyiekkl
 301 klkvnskksk vvrpwdlkyl gfsfyvkeek yeirvhgksi kefkkklkge
 351 tkrssgrsma yrlsrikqii tgwtnyygia nmksvagsld gwlrrrirmc
 401 iwkewkkikt khenlvklgl ntykaweyan trkgywrisn spilsmtltn
 451 krlkemglts iletynlkhq fc
```

Length in amino acids: 472
Molecular weight in kD: 54.66
Iso-electric point (IP): 9.82

Petrotoga mobilis SJ95 IEP3 Gene Sequence

```
   1 ttgaaagatg cgaagagaca cggaaaaagc atacaacttc gattagaagg
  51 tttcctacat gaagataagg gagagcctga aaataatgta gaagcgccta
 101 gtgtaacttc tacgtctgaa agaggaagaa acgatgataa aggatacagt
 151 gaagggatgc ttgaaaagat attatccaag gataatatga ataaagcata
 201 taagaaggta aaggccaaca aaggagcccc tggaatagac gggatggaag
 251 tagaagaact ctttgaatat ctcaaacaac atggagaaga attaaggcaa
 301 gagctccttg aaggaaggta caccccgaac ccggtaagga ggaaagagat
 351 accgaaaccg gatggaggga aaagactact aggcatacca acagcaatag
 401 acagagtaat ccaacaatcc atagcccagg aactgatacc gatttatgaa
 451 aagaaatttg tagataacag ctatggattt aggccattac gagatgcaaa
 501 acaagccata cggaaaagca aagaatacct aaacgaagga cacacgtggg
 551 tagtggacat agatttagaa cgatactttg acacagtcaa ccatgacaaa
 601 ctgatgagga taatatcaaa agacgtaaaa gatggcagag tgatatccct
 651 gataaggaaa tacctcaaga gtggggtaat ggtaaatggg gtagtaatag
 701 aaacagaaga agggactccg caagggggac cgctatcccc attactaagc
 751 aacataatgc tccacgaact tgacgtagaa ctaacgaaaa ggggacacaa
 801 gttttgcagg tatgcagacg attgcaacat atacgtgaaa agcgagaaat
 851 ccgcatatag ggtgatggaa agtataacaa aatacataga gaagaaatta
 901 aagttaaaag tgaacaggaa gaaaagcaaa gtagtcaagc ctttggacct
 951 caaatactta gggttctcct tctacgggaa agaagagcaa tacgaaatca
1001 gagtgcatga gaaatccatc aaagaattta aaaagaagtt aaaagaagaa
1051 acgaaaagaa gcagtggaag gagcatgaca tacaggctgt caaagataaa
1101 acaaataata acaggatgga taaactacta tggaatagca aacatgaaat
1151 cggcagcgga aagtcttgat ggatggctta gaaggagaat taggatgtgt
1201 atctggaagg aatggaagaa aatcaaaaca aggtataaaa acttagtcaa
1251 attaggacta aacacataca aagcatggga atatgcaaac acaagaaaag
1301 gccattggag gatctccaac agcccgatac tttcaagaac actaactaat
1351 aagcacctta aggaaattgg attaacttcg atcctggaaa catataacct
1401 aaaacatcaa ttctgttga
```

Petrotoga mobilis SJ95 IEP3 Sequence

```
   1 lkdakrhgks iqlrlegflh edkgepennv eapsvtstse rgrnddkgys
  51 egmlekilsk dnmnkaykkv kankgapgid gmeveelfey lkqhgeelrq
 101 ellegrytpn pvrrkeipkp dggkrllgip taidrviqqs iaqelipiye
 151 kkfvdnsygf rplrdakqai rkskeylneg htwvvdidle ryfdtvnhdk
 201 lmriiskdvk dgrvislirk ylksgvmvng vvieteegtp qgggplsplls
 251 nimlheldve ltkrghkfcr yaddcniyvk seksayrvme sitkyiekkl
 301 klkvnrkksk vvkpldlkyl gfsfygkeeq yeirvheksi kefkkklkee
 351 tkrssgrsmt yrlskikqii tgwinyygia nmksaaesld gwlrrrirmc
 401 iwkewkkikt ryknlvklgl ntykaweyan trkghwrisn spilsrtltn
 451 khlkeiglts iletynlkhq fc
```

Length in amino acids: 472
Molecular weight in kD: 54.66
Iso-electric point (IP): 9.82

Bacillus stearothermophilis trt DNA Sequence

```
   1 atggctttgt tggaacgcat cttagcgaga gacaacctca tcacggcgct
  51 caaacgggtc gaagccaacc aaggagcacc gggaatcgac ggagtatcaa
 101 ccgatcaact ccgtgattac atccgcgctc actggagcac gatccgcgcc
 151 caactcttgg cgggaaccta ccggccggcg cctgtccgca gggtcgaaat
 201 cccgaaaccg ggcggcggca cacggcagct aggcattccc accgtggtgg
 251 accggctgat ccaacaagcc attcttcaag aactcacacc cattttcgat
 301 ccagacttct ccccttccag cttcggattc cgtccgggcc gcaacgccca
 351 cgatgccgtg cggcaagcgc aaggctacat ccaggaagga tatcggtacg
 401 tggtcgacat ggacctggaa aagttctttg atcgggtcaa ccatgacatc
 451 ttgatgagtc gggtggcccg aaaagtcaag gataaacgcg tgctgaaact
 501 gatccgtgcc tacctgcaag ccggcgttat gatcgaaggg gtgaaggtgc
 551 agacggagga agggacgccg caaggcggcc ccctcagccc cctgctggcg
 601 aacatccttc tcgacgattt agacaaggaa ttggagaagc gaggattgaa
 651 attctgccgt tacgcagatg actgcaacat ctatgtgaaa agtctgcggg
 701 caggacaacg ggtgaaacaa agcatccaac ggttcttgga gaaaacgctc
 751 aaactcaaag taaacgagga gaaaagtgcg gtggaccgcc cgtggaaacg
 801 ggcctttctg gggtttagct tcacaccgga acgaaaagcg cgaatccggc
 851 tcgccccaag gtcgattcaa cgtctgaaac agcggattcg acagctgacc
 901 aacccaaact ggagcatatc gatgccagaa cgaattcatc gcgtcaatca
 951 atacgtcatg ggatggatcg ggtattttcg gctcgtcgaa accccgtctg
1001 tccttcagac catcgaagga tggattcgga ggaggcttcg actctgtcaa
1051 tggcttcaat ggaaacgggt cagaaccaga atccgtgagt taagagcgct
1101 ggggctgaaa gagacagcgg tgatggagat cgccaatacc cgaaaaggag
1151 cttggcgaac aacgaaaacg ccgcaactcc accaggccct gggcaagacc
1201 tactggaccg ctcaagggct caagagtttg acgcaacgat atttcgaact
1251 ccgtcaaggt tga
```

Bacillus stearothermophilis trt Protein Sequence

```
   1 MALLERILAR DNLITALKRV EANQGAPGID GVSTDQLRDY IRAHWSTIRA
  51 QLLAGTYRPA PVRRVEIPKP GGGTRQLGIP TVVDRLIQQA ILQELTPIFD
 101 PDFSPSSFGF RPGRNAHDAV RQAQGYIQEG YRYVVDMDLE KFFDRVNHDI
 151 LMSRVARKVK DKRVLKLIRA YLQAGVMIEG VKVQTEEGTP QGGPLSPLLA
 201 NILLDDLDKE LEKRGLKFCR YADDCNIYVK SLRAGQRVKQ SIQRFLEKTL
 251 KLKVNEEKSA VDRPWKRAFL GFSFTPERKA RIRLAPRSIQ RLKQRIRQLT
 301 NPNWSISMPE RIHRVNQYVM GWIGYFRLVE TPSVLQTIEG WIRRRLRLCQ
 351 WLQWKRVRTR IRELRALGLK ETAVMEIANT RKGAWRTTKT PQLHQALGKT
 401 YWTAQGLKSL TQRYFELRQG
```

Length in amino acids: 420
Molecular weight in kD: 48.65
Iso-electric point (IP): 10.54

New Zealand *Bacillus stearothermophilus* Strain 11 IEP Gene Sequence

```
   1 atggctttgt tggaacgcat cttagcgaga gacaacctca ccacggcgct
  51 caaacgggtc gaagcgaacc aaggagcacc gggaatcgac ggagtatcaa
 101 ccgatcaact ccgtgattcc atccgcgctc actggggcac gatccgcgcc
 151 caactcttgg cgggaaccta ccggccggcg cctgtccgca gggtcgaaat
 201 cccgaaaccc agcggcggca cacggcagct aggcattccc accgtggtgg
 251 accggctgat ccaacaagcc attcttcaag aactcacccc cattttcgat
 301 ccagacttct ccccgtccag cttcggattc cgtccgggcc gcaacgctca
 351 cgatgccgtg cggcaagcgc aaggctacat ccaggagggg tatcggtacg
 401 tggtcgacat ggacctggaa aagttctttg atcgggtcaa ccatgacatc
 451 ctgatgagtc gggtggcccg aaaagtcaag gataaacgcg tgctgaaact
 501 gatccgtgcc tacctgcaag ccggcgttat gatcgaaggg gtgaaggtgc
 551 agacggagga agggacgccg caaggcggtc ccctcagccc cctgctggcg
 601 aacatccttc tcgacgattt agacaaggag ttggagaagc gcggattgaa
 651 attctgccgc tacgcagatg acggcaacat ctatgtgaaa agtctgcggg
 701 cagggcaacg ggtgaaacaa agcatccaac ggttcttgga gaaaacgctc
 751 aaactcaaag taaacgagga gaaaagtgcg gtggaccgcc cgtggaaacg
 801 ggcctttcta gggtttagct tcacaccgga acgaaaagcg cgaatccggc
 851 ttgccccaag gtcgattcaa cgtctgaaac agcggattcg acagctgacg
 901 aacccaaact ggagcctatc gatgccagaa cgaattcatc gtgtccatca
 951 atacgtcatg ggatggatcg ggtattttcg gctcgtcgaa accccgtctg
1001 tccttcagac catcgaagga tggattcgga ggaggcttcg actctgtcaa
1051 tggcttcaat ggaaacgggt cagaaccaga attcgtgagt taagggcgct
1101 ggggttgaag gagacagcgg tgatggagat cgccaatacc cgaaaaggag
1151 cttggcgaac aacgaaaacg ccacaactcc accaagccat gggcaaggac
1201 tattggactg tccaaggact caagagtttg acgcaacgat atttcgaact
1251 ccgtcaaggt taa
```

New Zealand *Bacillus stearothermophilus* Strain 11 IEP Sequence

```
   1 mallerilar dnlttalkrv eanqgapgid gvstdqlrds irahwgtira
  51 qllagtyrpa pvrrveipkp sggtrqlgip tvvdrliqqa ilqeltpifd
 101 pdfspssfgf rpgrnahdav rqaqgyiqeg yryvvdmdle kffdrvnhdi
 151 lmsrvarkvk dkrvlklira ylqagvmieg vkvqteegtp qggplsplla
 201 nillddldke lekrglkfcr yaddgniyvk slragqrvkq siqrflektl
 251 klkvneeksa vdrpwkrafl gfsftperka rirlaprsiq rlkqrirqlt
 301 npnwslsmpe rihrvhqyvm gwigyfrlve tpsvlqtieg wirrrlrlcq
 351 wlqwkrvrtr irelralglk etavmeiant rkgawrttkt pqlhqamgkd
 401 ywtvqglksl tqryfelrqg
```

Length in amino acids: 420
Molecular weight in kD: 48.60
Iso-electric point (IP): 10.58

IV – Sequence Alignments

New Zealand strain *B. stearothermophilus* IEP compared to the *B. stearothermophilus* Trt protein

```
NZ              MALLERILARDNLTTALKRVEANQGAPGIDGVSTDQLRDSIRAHWGTIRAQLLAGTYRPA
B.st            MALLERILARDNLITALKRVEANQGAPGIDGVSTDQLRDYIRAHWSTIHAQLLAGTYRPA
                ************ ************************ ******.**:***********

NZ              PVRRVEIPKPSGGTRQLGIPTVVDRLIQQAILQELTPIFDPDFSPSSFGFRPGRNAHDAV
B.st            PVRRVEIPKPGGGTRQLGIPTVVDRLIQQAILQELTPIFDPDFSSSSFGFRPGRNAHDAV
                **********.*********************************.***************

NZ              RQAQGYIQEGYRYVVDMDLEKFFDRVNHDILMSRVARKVKDKRVLKLIRAYLQAGVMIEG
B.st            RQAQGYIQEGYRYVVDMDLEKFFDRVNHDILMSRVARKVKDKRVLKLIRAYLQAGVMIEG
                ************************************************************

NZ              VKVQTEEGTPQGGPLSPLLANILLDDLDKELEKRGLKFCRYADDGNIYVKSLRAGQRVKQ
B.st            VKVQTEEGTPQGGPLSPLLANILLDDLDKELEKRGLKFCRYADDCNIYVKSLRAGQRVKQ
                ****************************************** ***************

NZ              SIQRFLEKTLKLKVNEEKSAVDRPWKRAFLGFSFTPERKARIRLAPRSIQRLKQRIRQLT
B.st            SIQRFLEKTLKLKVNEEKSAVDRPWKRAFLGFSFTPERKARIRLAPRSIQRLKQRIRQLT
                ************************************************************

NZ              NPNWSLSMPERIHRVHQYVMGWIGYFRLVETPSVLQTIEGWIRRRLRLCQWLQWKRVRTR
B.st            NPNWSISMPERIHRVNQYVMGWIGYFRLVETPSVLQTIEGWIRRRLRLCQWLQWKRVRTR
                *****:*********:********************************************

NZ              IRELRALGLKETAVMEIANTRKGAWRTTKTPQLHQAMGKDYWTVQGLKSLTQRYFELRQG
B.st            IRELRALGLKETAVMEIANTRKGAWRTTKTPQLHQALGKTYWTAQGLKSLTQRYFELRQG
                ************************************:** ***.****************

NZ              ATGGCTTTGTTGGAACGCATCTTAGCGAGAGACAACCTCACCACGGCGCTCAAACGGGTC
B.st            ATGGCTTTGTTGGAACGCATCTTAGCGAGAGACAACCTCATCACGGCGCTCAAACGGGTC
                ************************************ ******************

NZ              GAAGCGAACCAAGGAGCACCGGGAATCGACGGAGTATCAACCGATCAACTCCGTGATTCC
B.st            GAAGCCAACCAAGGAGCACCGGGAATCGACGGAGTATCAACCGATCAACTCCGTGATTAC
                ***** ********************************************************

NZ              ATCCGCGCTCACTGGGGCACGATCCGCGCCCAACTCTTGGCGGGAACCTACCGGCCGGCG
B.st            ATCCGCGCTCACTGGAGCACGATCCACGCCCAACTCTTGGCGGGAACCTACCGGCCGGCG
                *************** ********* **********************************

NZ              CCTGTCCGCAGGGTCGAAATCCCGAAACCCAGCGGCGGCACACGGCAGCTAGGCATTCCC
B.st            CCTGTCCGCAGGGTCGAAATCCCGAAACCGGGCGGCGGCACACGGCAGCTAGGCATTCCC
                ***************************  ******************************

NZ              ACCGTGGTGGACCGGCTGATCCAACAAGCCATTCTTCAAGAACTCACCCCCATTTTCGAT
B.st            ACCGTGGTGGACCGGCTGATCCAACAAGCCATTCTTCAAGAACTCACACCCATTTTCGAT
                ********************************************* ************

NZ              CCAGACTTCTCCCCCGTCCAGCTTCGGATTCCGTCCGGGCCGCAACGCTCACGATGCCGTG
B.st            CCAGACTTCTCCTCTTCCAGCTTCGGATTCCGTCCGGGCCGCAACGCCCACGATGCCGTG
                ************ * ********************************* ************

NZ              CGGCAAGCGCAAGGCTACATCCAGGAGGGGTATCGGTACGTGGTCGACATGGACCTGGAA
B.st            CGGCAAGCGCAAGGCTACATCCAGGAAGGGTATCGGTACGTGGTCGACATGGACCTGGAA
                ************************* ********************************
```

```
NZ              AAGTTCTTTGATCGGGTCAACCATGACATCCTGATGAGTCGGGTGGCCCGAAAAGTCAAG
B.st            AAGTTCTTTGATCGGGTCAACCATGACATCTTGATGAGTCGGGTGGCCCGAAAAGTCAAG
                ****************************** ******************************

NZ              GATAAACGCGTGCTGAAACTGATCCGTGCCTACCTGCAAGCCGGCGTTATGATCGAAGGG
B.st            GATAAACGCGTGCTGAAACTGATCCGTGCCTACCTGCAAGCCGGCGTTATGATCGAAGGG
                ************************************************************

NZ              GTGAAGGTGCAGACGGAGGAAGGGACGCCGCAAGGCGGTCCCCTCAGCCCCCTGCTGGCG
B.st            GTGAAGGTGCAGACGGAGGAAGGGACGCCGCAAGGCGGCCCCCTCAGCCCCCTGCTGGCG
                ************************************ ********************

NZ              AACATCCTTCTCGACGATTTAGACAAGGAGTTGGAGAAGCGCGGATTGAAATTCTGCCGC
B.st            AACATCCTTCTCGACGATTTAGACAAGGAATTGGAGAAGCGAGGATTGAAATTCTGCCGT
                **************************** ************ *****************

NZ              TACGCAGATGACGGCAACATCTATGTGAAAAGTCTGCGGGCAGGGCAACGGGTGAAACAA
B.st            TACGCAGATGACTGCAACATCTATGTGAAAAGTCTGCGGGCAGGACAACGGGTGAAACAA
                ************ ****************************** ***************

NZ              AGCATCCAACGGTTCTTGGAGAAAACGCTCAAACTCAAAGTAAACGAGGAGAAAAGTGCG
B.st            AGCATCCAACGGTTCTTGGAGAAAACGCTCAAACTCAAAGTAAACGAGGAGAAAAGTGCG
                ************************************************************

NZ              GTGGACCGCCCGTGGAAACGGGCCTTTCTAGGGTTTAGCTTCACACCGGAACGAAAAGCG
B.st            GTGGACCGCCCGTGGAAACGGGCCTTTCTGGGGTTTAGCTTCACACCGGAACGAAAAGCG
                **************************** ******************************

NZ              CGAATCCGGCTTGCCCCAAGGTCGATTCAACGTCTGAAACAGCGGATTCGACAGCTGACG
B.st            CGAATCCGGCTCGCCCCAAGGTCGATTCAACGTCTGAAACAGCGGATTCGACAGCTGACC
                *********** ********************************************** 


NZ              AACCCAAACTGGAGCCTATCGATGCCAGAACGAATTCATCGTGTCCATCAATACGTCATG
B.st            AACCCAAACTGGAGCATATCGATGCCAGAACGAATTCATCGCGTCAATCAATACGTCATG
                *************** ************************ *** **************

NZ              GGATGGATCGGGTATTTTCGGCTCGTCGAAACCCCGTCTGTCCTTCAGACCATCGAAGGA
B.st            GGATGGATCGGGTATTTTCGGCTCGTCGAAACCCCGTCTGTCCTTCAGACCATCGAAGGA
                ************************************************************

NZ              TGGATTCGGAGGAGGCTTCGACTCTGTCAATGGCTTCAATGGAAACGGGTCAGAACCAGA
B.st            TGGATTCGGAGGAGGCTTCGACTCTGTCAATGGCTTCAATGGAAACGGGTCAGAACCAGA
                ************************************************************

NZ              ATTCGTGAGTTAAGGGCGCTGGGGTTGAAGGAGACAGCGGTGATGGAGATCGCCAATACC
B.st            ATCCGTGAGTTAAGAGCGCTGGGGCTGAAAGAGACAGCGGTGATGGAGATCGCCAATACC
                ** *********** ********* **** ******************************

NZ              CGAAAAGGAGCTTGGCGAACAACGAAAACGCCACAACTCCACCAAGCCATGGGCAAGGAC
B.st            CGAAAAGGAGCTTGGCGAACAACGAAAACGCCGCAACTCCACCAGGCCCTGGGCAAGACC
                ******************************** *********** *** ********  *

NZ              TATTGGACTGTCCAAGGACTCAAGAGTTTGACGCAACGATATTTCGAACTCCGTCAAGGT
B.st            TACTGGACCGCTCAAGGGCTCAAGAGTTTGACGCAACGATATTTCGAACTCCGTCAAGGT
                ** ***** *  ***** ******************************************

NZ              TAA
B.st            TGA
                * *
```

P. mobilis IEP DNA sequence alignment

```
PMIEP1_DNA      TTGAAAGATGCGAAGGGACACGGAAAAAGCATACAACTTCGATTAGAAGGTTTCCTACAT
PMIEP2_DNA      TTGAAAGATGCGAAGGGACACGGAAAAAGCATACAACTTCGATTAGAAGGTTTCCTACAT
PMIEP3_DNA      TTGAAAGATGCGAAGAGACACGGAAAAAGCATACAACTTCGATTAGAAGGTTTCCTACAT
                *************** ********************************************

PMIEP1_DNA      GAAGATAAGGGAGAGCCTGAAAATAATGTAGAAGCGCCTAGTATAACTTCTACGTCTGAA
PMIEP2_DNA      GAAGATAAGGGAGAGCCTGAAAATAATGTAGAAGCGCCTAGTATAACTTCTACGTCTGAA
PMIEP3_DNA      GAAGATAAGGGAGAGCCTGAAAATAATGTAGAAGCGCCTAGTGTAACTTCTACGTCTGAA
                ****************************************** *****************

PMIEP1_DNA      AGAGGAAGAAACGATGATAAGGGATGCAGTGAAGGGATGCTTGAAAAGATATTATCCAAG
PMIEP2_DNA      AGAGGAAGAAACGATGATAAGGGATGCAGTGAAGGGATGCTTGAAAAGATATTATCCAAG
PMIEP3_DNA      AGAGGAAGAAACGATGATAAAGGATACAGTGAAGGGATGCTTGAAAAGATATTATCCAAG
                ******************** **** ***********************************

PMIEP1_DNA      GATAATATGAATAAAGCATACAAGAAGGTAAAGGCCAACAAAGGAGCCCCTGGGATAGAC
PMIEP2_DNA      GATAATATGAATAAAGCATACAAGAAGGTAAAGGCCAACAAAGGAGCCCCTGGGATAGAC
PMIEP3_DNA      GATAATATGAATAAAGCATATAAGAAGGTAAAGGCCAACAAAGGAGCCCCTGGAATAGAC
                ******************** ********************************** ******

PMIEP1_DNA      GGGATGAAAGTAGAAGAACTCTTTGGATATCTCAGACAACACGGAGAAGAATTAAGGCAA
PMIEP2_DNA      GGGATGAAAGTAGAAGAACTCTTTGGATATCTCAGACAACACGGAGAAGAATTAAGGCAA
PMIEP3_DNA      GGGATGGAAGTAGAAGAACTCTTTGAATATCTCAAACAACATGGAGAAGAATTAAGGCAA
                ****** ****************** ******** ****** ******************

PMIEP1_DNA      GAGCTCCTTGAAGGAAGGTACACCCCGAAGTCGGTAAGGAGGAAAGAAATACCGAAACCG
PMIEP2_DNA      GAGCTCCTTGAAGGAAGGTACACCCCGAAGTCGGTAAGGAGGAAAGAAATACCGAAACCG
PMIEP3_DNA      GAGCTCCTTGAAGGAAGGTACACCCCGAACCCGGTAAGGAGGAAAGAGATACCGAAACCG
                ******************************  ***************** ************

PMIEP1_DNA      GATGGAGGGAAAAGACTACTAGGCATACCAACATCAATTGACAGAGTAATCCAGCAATCC
PMIEP2_DNA      GATGGAGGGAAAAGACTACTAGGCATACCAACATCAATTGACAGAGTAATCCAGCAATCC
PMIEP3_DNA      GATGGAGGGAAAAGACTACTAGGCATACCAACAGCAATAGACAGAGTAATCCAACAATCC
                ********************************** **** ************** ******

PMIEP1_DNA      ATAGCCCAGGTGTTGACACCTATTTACGAAAAGAAATTTGTAGATAACAGTTATGGATTC
PMIEP2_DNA      ATAGCCCAGGTGTTGACACCTATTTACGAAAAGAAATTTGTAGATAACAGTTATGGATTC
PMIEP3_DNA      ATAGCCCAGGAACTGATACCGATTTATGAAAAGAAATTTGTAGATAACAGCTATGGATTT
                **********   *** *** ***** ************************ ********

PMIEP1_DNA      AGGCCATTACGGGATGCAAAACAAGCCATACGAAAAAGCAAAGAATACCTAAACAAAGGG
PMIEP2_DNA      AGGCCATTACGGGATGCAAAACAAGCCATACGAAAAAGCAAAGAATACCTAAACAAAGGG
PMIEP3_DNA      AGGCCATTACGAGATGCAAAACAAGCCATACGGAAAAGCAAAGAATACCTAAACGAAGGA
                *********** ******************** ********************** ****

PMIEP1_DNA      CATACATGGGTAGTGGACATAGACTTAGAACGCTACTTTGACACAGTCAACCATGACAAA
PMIEP2_DNA      CATACATGGGTAGTGGACATAGACTTAGAACGCTACTTTGACACAGTCAACCATGACAAA
PMIEP3_DNA      CACACGTGGGTAGTGGACATAGATTTAGAACGATACTTTGACACAGTCAACCATGACAAA
                ** ** ***************** ******** ****************************

PMIEP1_DNA      CTGATGAGGATAATATCAAAAGACGTAAAAGATGGCAGAGTGATATCCCTGATAAGGAAA
PMIEP2_DNA      CTGATGAGGATAATATCAAAAGACGTAAAAGATGGCAGAGTGATATCCCTGATAAGGAAA
PMIEP3_DNA      CTGATGAGGATAATATCAAAAGACGTAAAAGATGGCAGAGTGATATCCCTGATAAGGAAA
                ************************************************************

PMIEP1_DNA      TACCTCAAGAGTGGGGTAATGGTAAATGGGGTAGTAATAGAAACTGAAGAAGGGACTCCA
PMIEP2_DNA      TACCTCAAGAGTGGGGTAATGGTAAATGGGGTAGTAATAGAAACTGAAGAAGGGACTCCA
PMIEP3_DNA      TACCTCAAGAGTGGGGTAATGGTAAATGGGGTAGTAATAGAAACAGAAGAAGGGACTCCG
                ****************************************** **************
```

```
PMIEP1_DNA      CAAGGGGGACCGCTATCCCCATTACTAAGCAACATAATGCTCCACGAACTTGACGTAGAA
PMIEP2_DNA      CAAGGGGGACCGCTATCCCCATTACTAAGCAACATAATGCTCCACGAACTTGACGTAGAA
PMIEP3_DNA      CAAGGGGGACCGCTATCCCCATTACTAAGCAACATAATGCTCCACGAACTTGACGTAGAA
                ************************************************************

PMIEP1_DNA      CTAACGAAAAGGGGACACAAGTTTTGCAGGTATGCAGACGATTGCAACATATACGTGAAA
PMIEP2_DNA      CTAACGAAAAGGGGACACAAGTTTTGCAGGTATGCAGACGATTGCAACATATACGTGAAA
PMIEP3_DNA      CTAACGAAAAGGGGACACAAGTTTTGCAGGTATGCAGACGATTGCAACATATACGTGAAA
                ************************************************************

PMIEP1_DNA      AGCGAGAAATCCGCATATAGGGTGATGGAAAGTATAACAAAATACATAGAGAAGAAATTA
PMIEP2_DNA      AGCGAGAAATCCGCATATAGGGTGATGGAAAGTATAACAAAATACATAGAGAAGAAATTA
PMIEP3_DNA      AGCGAGAAATCCGCATATAGGGTGATGGAAAGTATAACAAAATACATAGAGAAGAAATTA
                ************************************************************

PMIEP1_DNA      AAGCTAAAAGTGAATAGCAAGAAAAGCAAAGTAGTCAGACCTTGGGACCTCAAATACTTA
PMIEP2_DNA      AAGCTAAAAGTGAATAGCAAGAAAAGCAAAGTAGTCAGACCTTGGGACCTCAAATACTTA
PMIEP3_DNA      AAGTTAAAAGTGAACAGGAAGAAAAGCAAAGTAGTCAAGCCTTTGGACCTCAAATACTTA
                *** ********** ** ********************  **** ****************

PMIEP1_DNA      GGGTTCTCCTTCTACGTGAAAGAAGAGAAGTACGAAATTAGAGTCCATGGAAAATCCATT
PMIEP2_DNA      GGGTTCTCCTTCTACGTGAAAGAAGAGAAGTACGAAATTAGAGTCCATGGAAAATCCATT
PMIEP3_DNA      GGGTTCTCCTTCTACGGGAAAGAAGAGCAATACGAAATCAGAGTGCATGAGAAATCCATC
                ***************** ********** * ******** ***** ****  ********

PMIEP1_DNA      AAAGAATTCAAAAAGAAGTTAAAAGGAGAAACGAAGAGAAGCAGTGGAAGAAGCATGGCA
PMIEP2_DNA      AAAGAATTCAAAAAGAAGTTAAAAGGAGAAACGAAGAGAAGCAGTGGAAGAAGCATGGCA
PMIEP3_DNA      AAAGAATTTAAAAAGAAGTTAAAAGAAGAAACGAAAAGAAGCAGTGGAAGGAGCATGACA
                ******** ***************** ********* ************** ****** **

PMIEP1_DNA      TACAGGCTGTCAAGGATAAAACAAATAATAACAGGATGGACAAACTACTATGGAATAGCA
PMIEP2_DNA      TACAGGCTGTCAAGGATAAAACAAATAATAACAGGATGGACAAACTACTATGGAATAGCA
PMIEP3_DNA      TACAGGCTGTCAAAGATAAAACAAATAATAACAGGATGGATAAACTACTATGGAATAGCA
                ************* ************************* ********************

PMIEP1_DNA      AACATGAAATCGATAGCGGGAAGTCTTGATGGATGGCTTAGAAGGAGAATTAGGATGTGT
PMIEP2_DNA      AACATGAAATCGGTAGCGGGAAGTCTTGATGGATGGCTCAGAAGGAGAATTAGGATGTGT
PMIEP3_DNA      AACATGAAATCGGCAGCGGAAAGTCTTGATGGATGGCTTAGAAGGAGAATTAGGATGTGT
                ************  ***** ****************** ***********************

PMIEP1_DNA      ATCTGGAAGGAATGGAAGAAAATCAAAACAAGGTATAAAAACTTAGTCAAATTAGGGTTA
PMIEP2_DNA      ATCTGGAAGGAATGGAAGAAAATCAAAACAAAACATGAAAACTTAGTCAAATTAGGACTA
PMIEP3_DNA      ATCTGGAAGGAATGGAAGAAAATCAAAACAAGGTATAAAAACTTAGTCAAATTAGGACTA
                ********************************   ** ******************  **

PMIEP1_DNA      AACATCTACAAAGCATGGGAATATGCAAATACAAGGAAAGGCTATTGGAGAATCTCCAAC
PMIEP2_DNA      AACACCTACAAAGCATGGGAATATGCAAACACAAGAAAAGGCTATTGGAGGATCTCCAAC
PMIEP3_DNA      AACACATACAAAGCATGGGAATATGCAAACACAAGAAAAGGCCATTGGAGGATCTCCAAC
                ****  ********************** ***** ****** ******* *********

PMIEP1_DNA      AGCCCCATACTTTCAAGAACACTAACTAATAAACACCTTAAGAAAATGGGGTTAACTTCG
PMIEP2_DNA      AGCCCGATACTTTCAATGACACTAACTAATAAACGCCTTAAGGAAATGGGGTTAACTTCG
PMIEP3_DNA      AGCCCGATACTTTCAAGAACACTAACTAATAAGCACCTTAAGGAAATTGGATTAACTTCG
                ***** **********  ************** * ******* **** ** *********

PMIEP1_DNA      ATCCTGGAAACGTATAACCTAAAGCATGAATTCTGTTGA
PMIEP2_DNA      ATCCTGGAAACGTATAACCTAAAACATCAATTCTGTTGA
PMIEP3_DNA      ATCCTGGAAACATATAACCTAAAACATCAATTCTGTTGA
                *********** *********** *** ***********
```

P. mobilis IEP amino acid sequence alignment

```
PMIEP1_PROTEIN      LKDAKGHGKSIQLRLEGFLHEDKGEPENNVEAPSITSTSERGRNDDKGCSEGMLEKILSK
PMIEP2_PROTEIN      LKDAKGHGKSIQLRLEGFLHEDKGEPENNVEAPSITSTSERGRNDDKGCSEGMLEKILSK
PMIEP3_PROTEIN      LKDAKRHGKSIQLRLEGFLHEDKGEPENNVEAPSVTSTSERGRNDDKGYSEGMLEKILSK
                    ***** ****************************:************* ***********

PMIEP1_PROTEIN      DNMNKAYKKVKANKGAPGIDGMKVEELFGYLRQHGEELRQELLEGRYTPKSVRRKEIPKP
PMIEP2_PROTEIN      DNMNKAYKKVKANKGAPGIDGMKVEELFGYLRQHGEELRQELLEGRYTPKSVRRKEIPKP
PMIEP3_PROTEIN      DNMNKAYKKVKANKGAPGIDGMEVEELFEYLKQHGEELRQELLEGRYTPNPVRRKEIPKP
                    **********************:***** **:*****************:.*********

PMIEP1_PROTEIN      DGGKRLLGIPTSIDRVIQQSIAQVLTPIYEKKFVDNSYGFRPLRDAKQAIRKSKEYLNKG
PMIEP2_PROTEIN      DGGKRLLGIPTSIDRVIQQSIAQVLTPIYEKKFVDNSYGFRPLRDAKQAIRKSKEYLNKG
PMIEP3_PROTEIN      DGGKRLLGIPTAIDRVIQQSIAQELIPIYEKKFVDNSYGFRPLRDAKQAIRKSKEYLNEG
                    ***********:*********** * *********************************:*

PMIEP1_PROTEIN      HTWVVDIDLERYFDTVNHDKLMRIISKDVKDGRVISLIRKYLKSGVMVNGVVIETEEGTP
PMIEP2_PROTEIN      HTWVVDIDLERYFDTVNHDKLMRIISKDVKDGRVISLIRKYLKSGVMVNGVVIETEEGTP
PMIEP3_PROTEIN      HTWVVDIDLERYFDTVNHDKLMRIISKDVKDGRVISLIRKYLKSGVMVNGVVIETEEGTP
                    ************************************************************

PMIEP1_PROTEIN      QGGPLSPLLSNIMLHELDVELTKRGHKFCRYADDCNIYVKSEKSAYRVMESITKYIEKKL
PMIEP2_PROTEIN      QGGPLSPLLSNIMLHELDVELTKRGHKFCRYADDCNIYVKSEKSAYRVMESITKYIEKKL
PMIEP3_PROTEIN      QGGPLSPLLSNIMLHELDVELTKRGHKFCRYADDCNIYVKSEKSAYRVMESITKYIEKKL
                    ************************************************************

PMIEP1_PROTEIN      KLKVNSKKSKVVRPWDLKYLGFSFYVKEEKYEIRVHGKSIKEFKKKLKGETKRSSGRSMA
PMIEP2_PROTEIN      KLKVNSKKSKVVRPWDLKYLGFSFYVKEEKYEIRVHGKSIKEFKKKLKGETKRSSGRSMA
PMIEP3_PROTEIN      KLKVNRKKSKVVKPLDLKYLGFSFYGKEEQYEIRVHEKSIKEFKKKLKEETKRSSGRSMT
                    ***** ******:* ********** ***:****** *********** **********:

PMIEP1_PROTEIN      YRLSRIKQIITGWTNYYGIANMKSIAGSLDGWLRRRIRMCIWKEWKKIKTRYKNLVKLGL
PMIEP2_PROTEIN      YRLSRIKQIITGWTNYYGIANMKSVAGSLDGWLRRRIRMCIWKEWKKIKTKHENLVKLGL
PMIEP3_PROTEIN      YRLSKIKQIITGWINYYGIANMKSAAESLDGWLRRRIRMCIWKEWKKIKTRYKNLVKLGL
                    ****:******** ********** * ***********************:::*******

PMIEP1_PROTEIN      NIYKAWEYANTRKGYWRISNSPILSRTLTNKHLKKMGLTSILETYNLKHEFC
PMIEP2_PROTEIN      NTYKAWEYANTRKGYWRISNSPILSMTLTNKRLKEMGLTSILETYNLKHQFC
PMIEP3_PROTEIN      NTYKAWEYANTRKGHWRISNSPILSRTLTNKHLKEIGLTSILETYNLKHQFC
                    * ************:********** *****:**::**************:**
```